Behaviour Research and Therapy 192 (2025) 104804

Contents lists available at ScienceDirect

Behaviour Research and Therapy

journal homepage: www.elsevier.com/locate/brat

Affective and computational determinants of threat extinction biases

Yoann Stussi [a,b,*]

[a] Swiss Center for Affective Sciences, Campus Biotech, University of Geneva, Geneva, Switzerland
[b] Department of Psychology, FPSE, University of Geneva, Geneva, Switzerland

ARTICLE INFO

Keywords:
Emotion
Learning
Pavlovian conditioning
Extinction
Computational modeling
Threat
Reward

ABSTRACT

Pavlovian threat acquisition and extinction are fundamental processes by which individuals learn about threat and safety in their environment. Research has shown that humans learn more rapidly and persistently to associate threatening and—somewhat counterintuitively—positive rewarding stimuli with aversive events, supporting predictions derived from appraisal theories of emotion (Stussi et al., 2018; Stussi, Pourtois et al., 2021). Here, the present study aimed to provide a confirmatory analysis of these findings and further characterize their algorithmic bases. Data from the four original experiments ($N = 247$) using a differential Pavlovian threat conditioning paradigm were combined and reanalyzed. In this paradigm, threat-relevant (angry faces, snakes), positive-relevant (baby faces, happy faces, erotic images), and neutral (neutral faces, colored squares) stimuli were used as conditioned stimuli, and skin conductance response was measured as an index of learning. Computational modeling was applied to identify signatures of learning biases in Pavlovian threat acquisition and extinction. An expanded model comparison indicated that a reinforcement-learning model differentiating between excitatory (learning from reinforcement) and inhibitory (learning from the absence of reinforcement) learning best explained the observed data. Although no evidence for differences in excitatory learning rates was found between stimulus categories, both threat- and positive-relevant stimuli exhibited a lower inhibitory learning rate compared to neutral stimuli, contributing to the persistence of the conditioned response during extinction. These results confirm the robustness of the original findings and further validate the appraisal-based approach, thereby informing the affective and computational determinants of Pavlovian threat extinction biases and their translational relevance.

1. Introduction

Learning about threat and safety is pivotal for navigating fast-changing environments and helps organisms produce adaptive behaviors that promote survival (LeDoux & Daw, 2018; Levy & Schiller, 2021). Pavlovian conditioning processes, such as threat acquisition and extinction, are central to this learning. Pavlovian threat conditioning serves as a laboratory model for understanding the mechanisms underlying how defensive responses are learned and the etiology and maintenance of anxiety- and stress-related disorders (Beckers et al., 2023; Cooper et al., 2024; Zinbarg et al., 2022). Pavlovian threat acquisition models how multifaceted defensive responses are acquired through single or repeated contingent pairing between a cue (the conditioned stimulus) and an aversive event (the unconditioned stimulus; Betti et al., 2024; Lonsdorf et al., 2017; Ojala & Bach, 2020; Pavlov, 1927; Phelps, 2006; Rescorla, 1988). On the other hand, Pavlovian threat extinction models the gradual weakening or persistence of conditioned defensive responses when the conditioned stimulus is subsequently presented in safe contexts (i.e., in the absence of the unconditioned stimulus; Dunsmoor et al., 2015; Laing et al., 2025). Extinction (or safety) learning is notably a core principle of exposure therapy that is leveraged to reduce maladaptive defensive responses (Craske et al., 2014, 2022). Elucidating the mechanisms that modulate how conditioned defensive responses are acquired and extinguished thus holds high translational value across basic and clinical research, as well as clinical practice (Kredlow et al., 2018; Milad & Quirk, 2012; Zuj & Norrholm, 2019).

In this context, a critical notion is that some Pavlovian associations are more easily learned and maintained than others depending on the nature of the conditioned stimulus (Garcia & Koelling, 1966; Hamm et al., 1989; Seligman, 1970; Öhman & Mineka, 2001). Empirical evidence has shown that conditioned defensive responses to evolutionarily threat-relevant stimuli—such as snakes, angry faces, and outgroup

This article is part of a special issue entitled: Innovations and Advances in Human Fear Extinction.

* Campus Biotech, CISA-University of Geneva, Chemin des Mines 9, CH-1202, Geneva, Switzerland.
E-mail address: yoann.stussi@unige.ch.

https://doi.org/10.1016/j.brat.2025.104804
Received 30 November 2024; Received in revised form 28 March 2025; Accepted 12 June 2025
Available online 14 June 2025

faces—are more rapidly acquired (Ho & Lipp, 2014; Öhman et al., 1975) and more resistant to extinction (e.g., Olsson et al., 2005; Öhman & Dimberg, 1978; Öhman et al., 1976) than to threat-irrelevant stimuli—such as flowers, neutral faces, and ingroup faces (but see Åhs et al., 2018; McNally, 1987, for reviews indicating that these effects are not always replicated). These findings have generally been interpreted as supporting the preparedness (Seligman, 1970, 1971) and fear module (Öhman & Mineka, 2001) theories. According to these theories, organisms are biologically predisposed to associate stimuli that threatened the species' survival through evolution with aversive events, with the key argument that such highly "prepared" learning is at the origin of specific phobias (Seligman, 1971).

Departing from this threat-selectivity perspective, an alternative framework originating from appraisal theories of emotion (Moors et al., 2013; Scherer & Moors, 2019; Yeo & Ong, 2024) suggests that biases in Pavlovian threat acquisition and extinction are driven by affective relevance—a key appraisal process involved in emotion elicitation—rather than a threat-specific mechanism (Stussi et al., 2015, 2018, 2019; Stussi, Pourtois, et al., 2021). Affective relevance appraisal is conceptualized as a rapid and flexible process whereby the individual establishes whether a stimulus is likely to impact their major concerns—such as their goals, needs, values, or well-being (Frijda, 1986; Pool et al., 2016; Sander et al., 2003, 2018; Stussi et al., 2018; Stussi & Pool, 2022). The affective relevance model proposes that evolutionarily threat-relevant stimuli benefit from faster and more persistent Pavlovian threat conditioning because they are likely to be appraised as highly relevant to survival. Nonetheless, it suggests that such Pavlovian learning biases occur for stimuli appraised as affectively relevant to the individual's concerns beyond their valence and threat value. While this may appear counterintuitive, a key prediction is that positive stimuli with high affective relevance should also be preferentially associated with an aversive outcome during Pavlovian threat conditioning, similar to threat-relevant stimuli. A series of experiments directly tested this hypothesis and showed that both threat-relevant and positive-relevant stimuli (e.g., baby faces, happy faces, erotic stimuli) are more readily and persistently associated with an aversive unconditioned stimulus compared to neutral stimuli (Stussi et al., 2018; Stussi, Pourtois, et al., 2021; see also Ney et al., 2022), thereby supporting the affective relevance model.

Preliminary evidence at the algorithmic level suggested that a reinforcement-learning model differentiating excitatory (learning from reinforcement) and inhibitory (learning from the absence of reinforcement) learning best explained these effects. Specifically, threat- and positive-relevant stimuli were associated with lower inhibitory learning rates compared to neutral stimuli (Stussi et al., 2018; Stussi, Pourtois, et al., 2021). These reduced inhibitory learning rates diminished the impact of negative prediction errors (i.e., when the aversive outcome is omitted or less than predicted) on associative strength, contributing to the persistence of the conditioned defensive responses to threat- and positive-relevant stimuli. By contrast, no differences in excitatory learning were found across the conditioned stimulus categories. These results partly reflect core characteristics of anxiety- and stress-related disorders, such as increased defensive responses to safety signals and the persistence of these responses in the absence of threat (Abend et al., 2022; Duits et al., 2015; Homan et al., 2019; Kausche et al., 2025). Consequently, the robust and reliable identification and characterization of the affective and computational determinants of learning biases in Pavlovian threat acquisition and extinction are of high relevance for advancing both theoretical and clinical knowledge.

In this perspective, and in line with efforts to foster transparency, replicability, reproducibility, and credibility in human threat conditioning research (Bach et al., 2023; Cooper et al., 2023; Ehlers & Lonsdorf, 2022; Lonsdorf et al., 2017, 2019, 2022; Ney et al., 2018), the present study sought to provide a large-scale confirmatory analysis of the results reported by Stussi et al. (Stussi et al., 2018; Stussi, Pourtois, et al., 2021). To that end, data from the original studies (4 experiments, $N = 247$) were combined to increase statistical power. A reanalysis of skin conductance response (SCR) data, used as an index of the conditioned defensive response (CR), was conducted to evaluate the robustness of the original findings. This reanalysis employed a different analytical specification, leveraging trial-by-trial information (i.e., mixed-effects modeling) instead of aggregate difference scores across conditioned stimulus categories (see Ney et al., 2018). Importantly, the original computational analyses were expanded by considering a latent cause model in addition to reinforcement-learning models. Latent cause models of associative learning suggest that distinct learning dynamics in threat acquisition and extinction stem from the attribution of acquisition and extinction trials to separate latent causes (Gershman & Hartley, 2015). This alternative computational framework can account for the co-occurrence of faster and more persistent threat conditioning to affectively relevant stimuli, thus providing a more stringent comparison with the dual-learning-rates approach. In doing so, this work aimed to rigorously assess (a) the appraisal-based predictions that both threat-relevant and positive-relevant stimuli are more rapidly and persistently associated with an aversive event than neutral stimuli during Pavlovian threat conditioning, and (b) whether higher excitatory and lower inhibitory learning rates are robust computational signatures of these learning biases.

2. Method

2.1. Data and participants

Data from the four experiments of the original studies (Stussi et al., 2018; Stussi, Pourtois, et al., 2021) were compiled in the current reanalysis. The data from Experiments 1 to 3 were taken from Stussi et al. (2018) and the data from Experiment 4 were extracted from Stussi, Pourtois, et al. (2021). The experiments were approved by the Faculty of Psychology and Educational Sciences ethics committee at the University of Geneva or by the Regional Ethics Committee in Geneva (2016–01009).

A total of 312 volunteers were recruited ($n = 52$ in Experiment 1, $n = 88$ in Experiment 2, $n = 55$ men in Experiment 3, and $n = 117$ in Experiment 4). Based on the exclusion criteria of the original experiments (Stussi et al., 2018; Stussi, Pourtois, et al., 2021; see also Olsson et al., 2005), 65 participants were excluded from the analyses because of technical problems ($n = 19$), for displaying virtually no SCRs to the conditioned and unconditioned stimuli ($n = 16$), for failing to acquire a CR to at least one of the reinforced conditioned stimuli (i.e., the mean difference in SCR between the reinforced and unreinforced conditioned stimuli during acquisition was equal to or below 0 for all the stimulus categories; $n = 26$), or for withdrawing from the experiment early ($n = 4$).

The final sample included 247 participants (162 women, 85 men) aged between 18 and 52 years old ($M_{age} = 22.94$, $SD_{age} = 4.93$). The data from three female participants were further excluded from the computational analyses because of a lack of SCR to a specific stimulus category preventing the estimation of free parameters. A sensitivity power analysis conducted with G*Power 3 (Faul et al., 2007) indicated that the current sample size allowed for detecting a smallest population effect size of Cohen's $d_z = 0.16$–0.21 with a power ranging between 80 and 95% using a one-tailed paired-sample t-test.

2.2. Stimuli and materials

2.2.1. Conditioned stimuli

Six conditioned stimuli (CSs) divided in three categories were used in each experiment: two threat-relevant stimuli, two positive-relevant stimuli, and two neutral stimuli. In Experiments 1 and 2, the threat-relevant stimuli consisted of two male angry faces, the positive-relevant stimuli of two male baby faces, and the neutral stimuli of two male neutral faces. The angry (model numbers 23 and 46) and neutral (model numbers 15 and 25) faces were taken from the Radboud Faces Database (Langner et al., 2010) and the baby faces were selected from a set of infant faces (Van Duuren et al., 2003). In Experiment 3, the two most disliked snake images among a set of 12 snake images from the International Affective Picture System (IAPS numbers 1022, 1026, 1033, 1040, 1050, 1051, 1052, 1070, 1090, 1113, 1114, and 1120; Lang et al., 2008) served as threat-relevant stimuli. The two most liked erotic images among a set of 24 images (12 images of nude or partially nude men or women each; Stussi, Sennwald, et al., 2021) were used as positive-relevant stimuli. Finally, the two colored squares rated as the most neutral among 12 colored squares served as neutral stimuli. All CSs in each category were selected individually for each participant. In Experiment 4, two male angry (model numbers AM10ANS and AM29ANS), two male happy (AM07HAS and AM22HAS), and two male neutral (AM11NES and AM31NES) faces taken from the Karolinska Directed Emotional Faces (Lundqvist et al., 1998) were used as threat-relevant, positive-relevant, and neutral stimuli, respectively. In all the experiments, the stimuli were presented using MATLAB (The MathWorks Inc., Natick, MA; RRID:SCR_001622) with the Psychophysics Toolbox extensions (Brainard, 1997; Pelli, 1997; RRID:SCR_002881).

Independent (Experiments 1 and 2) or within-sample (Experiments 3 and 4) subjective ratings confirmed that the threat-relevant stimuli were evaluated as unpleasant, the positive-relevant stimuli as pleasant, and the neutral stimuli as relatively neutral on average (see Stussi et al., 2018; Stussi, Pourtois, et al., 2021). Each stimulus from the three stimulus categories served both as a reinforced stimulus (CS+) and as an unreinforced stimulus (CS-), counterbalanced across participants.

2.2.2. *Unconditioned stimulus*

The unconditioned stimulus (US) consisted of a mild electric stimulation (200-ms in Experiments 1, 3, and 4, and 10-ms in Experiment 2) delivered to the participants' right or dominant forearm through a Grass SD9 stimulator (50 pulses/s; Grass Medical Instruments, West Warwick, RI) charged with a stabilized current or a unipolar pulse electric stimulator (STM200; BIOPAC Systems Inc., Goleta, CA). The electric stimulation intensity was determined for each participant using a calibration procedure (see section 2.3.1).

2.2.3. *Skin conductance apparatus*

The SCR was measured with two pre-gelled disposable Ag-AgCl electrodes (11-mm contact diameter; Experiment 1) or two Ag-AgCl electrodes (6-mm contact diameter; Experiments 2 to 4) with 0.5 % NaCl electrolyte gel. The electrodes were attached to the distal phalanges of the second and third digits of the participants' left or nondominant hand. Skin conductance was measured exosomatically with a direct current and a constant voltage of 0.5 V. The skin conductance signal was continuously recorded during the Pavlovian threat conditioning procedure with a sampling rate of 1000 Hz through a BIOPAC MP150 system (Santa Barbara, CA). The SCR data were analyzed offline with AcqKnowledge software (Version 4.2 or 4.4; BIOPAC Systems Inc., Goleta, CA; RRID:SCR_014279).

2.3. *Procedure*

Upon arrival at the laboratory, participants were asked to wash their hand with warm water and were seated in front of a computer monitor in a quiet room. They were next informed about the general layout of the experiment and provided written informed consent. The electrodes for measuring SCR and delivering the electric stimulation were attached to participants before the start of a US calibration procedure, which was followed by the differential Pavlovian threat conditioning procedure. These two procedures were identical across all the experiments, apart from the CSs used. After conditioning, participants provided subjective ratings of CS-US contingency and CS pleasantness (see supplementary materials; arousal and relevance ratings were additionally collected in Experiment 4).

In Experiment 3, participants rated the snake, erotic, and colored squares images according to their liking and felt arousal to select the two images from each category that were used as CSs (Stussi et al., 2018). In Experiment 4, participants performed a Go/No-go Association Task (that did not include the CSs) and rated the CSs as a function of their pleasantness, arousal, and relevance before the US calibration procedure (Stussi, Pourtois, et al., 2021). The differential Pavlovian threat conditioning procedure that was included in each experiment is described in detail below.

2.3.1. *Unconditioned stimulus calibration*

A work-up procedure was conducted to individually calibrate the electric stimulation intensity. The stimulation intensity started at 20 V and was increased (or decreased upon participant's request) in steps of 5 V until the participant reported the stimulation as "uncomfortable but not painful" (Lonsdorf et al., 2017), with a maximum of 50 V. The selected stimulation intensity ($M_{exp1} = 36.75$ V, $SD_{exp1} = 8.03$; $M_{exp2} = 34.75$ V, $SD_{exp2} = 7.59$; $M_{exp3} = 29.75$ V, $SD_{exp3} = 7.34$; $M_{exp4} = 34.55$ V, $SD_{exp4} = 7.57$) was used throughout the experiment.

2.3.2. *Differential Pavlovian threat conditioning*

Before conditioning, participants were instructed that they will have to carefully watch visual stimuli presented on a computer screen and that a mild electric stimulation will be delivered on some of the trials. They were next given the following instructions: "The task will be essentially passive, but you should nonetheless try to notice a potential pattern between the presentation of the visual stimuli and the delivery of an electric stimulation." Participants were however not informed on the specific contingencies between the CSs and the US. The differential Pavlovian threat conditioning included three contiguous phases: habituation, acquisition, and extinction. In the habituation phase, each of the six CSs were presented twice without reinforcement. During acquisition, each CS was presented seven times. This phase always started with a reinforced CS+ trial. Each CS+ was paired with the US following a partial reinforcement schedule. Five of the seven CS+ presentations coterminated with the US delivery, whereas the CS- from each stimulus category was never associated with the US. The use of a partial reinforcement schedule aimed to potentiate resistance to extinction of the CR, thus optimizing the investigation of differences in extinction across the three stimulus categories. Because the CSs+ became predictive of the US only after their first association therewith, the first acquisition trial for each CS was omitted from the SCR analyses. The extinction phase consisted of six unreinforced presentations of each CS. During all the conditioning phases, the CS was presented for 6 s with an intertrial interval ranging between 12 and 15 s during which a fixation cross was displayed onscreen. The CSs' presentation order was pseudorandomized into eight different orders to systematically counterbalance the associations between the stimulus identities and CS type (CS+ vs. CS-) across the three stimulus categories (threat-relevant vs. positive-relevant vs. neutral).

2.4. *Skin conductance response scoring*

The SCR scoring procedure was based on a prior study using a similar within-participants differential threat conditioning paradigm that included multiple CS categories (Olsson et al., 2005). This study reported medium-to-large effect sizes for the CS+/CS- differentiation during acquisition ($d_z = 0.62$–1.45). SCR was scored on each trial as the trough-to-peak amplitude difference in skin conductance of the largest response starting in the 0.5–4.5-s temporal window following CS onset. The minimal response criterion was 0.02 μS. Responses below this criterion were scored as zero and remained in the analyses. Trough-to-peak SCR scoring was employed because it is one of the most prevalent and validated SCR quantification approach in Pavlovian threat conditioning (Kuhn et al., 2022). It has been shown to discriminate between CS+ and CS- without any statistically significant difference in precision compared to model-based approaches (Kuhn et al., 2022). The scoring window was wider than standard recommendations (1–4 s; e.g., Society for Psychophysiological Research Ad Hoc Committee on Electrodermal Measures, 2012) to maximize the detection of event-related SCRs as is often done in human threat conditioning research (e.g., Olsson et al., 2005; Starita et al., 2019, 2023; Stussi et al., 2015, 2019; Zhang et al., 2016).

The SCR data was preprocessed before analysis using a low-pass filter (Blackman 92 dB, 1 Hz). SCRs were detected automatically with the AcqKnowledge software and manually checked for artifacts and response (mis)detection. Trials containing artifacts impacting the scoring of event-related SCRs (1.78% in Experiment 1, 0.003% in Experiment 2, 0.005% in Experiment 3, and 0.17% in Experiment 4) were omitted from the analyses. The raw SCR scores were scaled according to each participant's mean unconditioned response (UR), and square-root-transformed to normalize the distributions. The UR was scored as the trough-to-peak amplitude difference in skin conductance of the largest response starting in the 0.5–4.5-s temporal window following US delivery, and the mean UR was calculated across all valid US trials for each participant.

2.5. Computational modeling

Computational models of associative learning were used to test at the algorithmic level how threat-relevant and positive-relevant stimuli influence Pavlovian threat acquisition and extinction relative to neutral stimuli. Specifically, this analysis examined how learning parameters formalizing the latent cognitive processes that drive Pavlovian learning on a trial-by-trial basis varied as a function of the stimulus categories.

2.5.1. Model space

The candidate models consisted of (1) the Rescorla-Wagner model (Rescorla & Wagner, 1972), (2) a variant of the Rescorla-Wagner model incorporating dual learning rates for excitatory and inhibitory learning (Niv et al., 2012; Starita et al., 2023; Stussi et al., 2018; Stussi, Pourtois, et al., 2021), (3) the hybrid model combining the Pearce-Hall associability mechanism with the Rescorla-Wagner model (Homan et al., 2019; Le Pelley, 2004; Li et al., 2011), (4) a variant of the hybrid model with dual (static) learning rates for excitatory and inhibitory learning, and (5) the latent cause model proposed by Gershman and colleagues (Gershman & Hartley, 2015; Gershman & Niv, 2012).

Rescorla-Wagner model. The Rescorla-Wagner (Rescorla & Wagner, 1972) is a standard model of associative learning. According to this model, associative learning occurs when there is a discrepancy between the expected and the actual outcome (i.e., a prediction error). Formally, the Rescorla-Wagner model posits that the expected value V of a given CS is updated based on the sum of the current expected value V and the prediction error between the expected value V and the outcome R at trial t, weighted by a constant learning rate η:

$$V_{t+1} = V_t + \eta \times (R_t - V_t)$$

where the learning rate η is a free parameter within the range [0, 1]. If the US was delivered on trial t, $R_t = 1$, else $R_t = 0$.

Rescorla-Wagner model with dual learning rates. This model is an adaptation of the Rescorla-Wagner model that differentiates excitatory from inhibitory learning. It incorporates different learning rates for positive (i.e., when the outcome is unexpectedly delivered or more than predicted; excitatory learning) and negative (i.e., when the outcome is unexpectedly omitted or less than predicted; inhibitory learning) errors instead of a single learning rate. Excitatory and inhibitory learning rates modulate the extent to which positive and negative prediction errors are integrated in the computation of the updated CS expected value, respectively (Niv & Schoenbaum, 2008). Formally, this model posits that the expected value V of a given CS is updated based on the sum of the current expected value V and the prediction error between the expected value V and the outcome R at trial t, weighted by distinct learning rates for positive and negative prediction errors:

$$V_{t+1} = V_t + \begin{cases} \eta^{+} \times (R_t - V_t) \; if \; R_t - V_t > 0 \\ \eta^{-} \times (R_t - V_t) \; if \; R_t - V_t \leq 0 \end{cases}$$

where the excitatory learning rate η^{+} and the inhibitory learning rate η^{-} are free parameters within the range [0, 1]. If the US was delivered on trial t, $R_t = 1$, else $R_t = 0$. In dissociating excitatory and inhibitory learning as a function of stimulus contingencies, this model parsimoniously accounts for how specific stimulus categories can accelerate acquisition (through the excitatory learning rate) and enhance resistance to extinction (through the inhibitory learning rate).

Hybrid model. The hybrid model (Le Pelley, 2004; Li et al., 2011) is a combination of the Rescorla-Wagner model and the Pearce-Hall model (Pearce & Hall, 1980). It maintains the basic assumption that learning is directly driven by prediction errors, while incorporating the Pearce-Hall associability mechanism. Associability acts as a dynamic learning rate that determines the extent to which prediction errors are weighted into the update of the CS expected value. The CS associability is modulated on a trial-by-trial basis as a function of unsigned past prediction errors. Specifically, the CS associability decreases when the CS accurately and reliably predicts the outcome, whereas it increases when the CS is an unreliable predictor of the outcome. In the hybrid model, the expected value V and associability κ of a given CS are updated as follows:

$$V_{t+1} = V_t + \eta \times \kappa_t \times (R_t - V_t)$$

$$\kappa_{t+1} = \gamma \times |R_t - V_t| + (1 - \gamma) \times \kappa_t$$

where the initial associability κ_0, the static learning rate η, and the associability weight γ are free parameters within the range [0, 1]. If the US was delivered on trial t, $R_t = 1$, else $R_t = 0$.

Hybrid model with dual learning rates. A variant of the hybrid model implementing dual static learning rates for positive and negative prediction errors was tested. In this modified version of the hybrid model, the expected value V and associability κ of a given CS are updated as follows:

$$V_{t+1} = V_t + \begin{cases} \eta^{+} \times \kappa_t \times (R_t - V_t) \; if \; R_t - V_t > 0 \\ \eta^{-} \times \kappa_t \times (R_t - V_t) \; if \; R_t - V_t \leq 0 \end{cases}$$

$$\kappa_{t+1} = \gamma \times |R_t - V_t| + (1 - \gamma) \times \kappa_t$$

where the initial associability κ_0, the static excitatory learning rate η^{+}, the static inhibitory learning rate η^{-}, and the associability weight γ are free parameters within the range [0, 1]. If the US was delivered on trial t, $R_t = 1$, else $R_t = 0$.

Latent cause model. Latent cause models of Pavlovian conditioning (Gershman et al., 2010; Gershman & Niv, 2012) suggest that individuals learn the relationship between the CS and the US by inferring hidden or "latent" causes responsible for their co-occurrence. The CR is driven by predictions about which latent cause is likely active on a given trial and whether the US is expected in that context (Gershman et al., 2010; Gershman & Niv, 2012; see also Dunsmoor et al., 2015). According to this class of models, trials that share a common pattern of observable stimuli are likely to be clustered together and represented by a single latent cause (e.g., an "acquisition" cause when the CS+ and US are contingently paired). By contrast, different patterns of observable stimuli or discrepancies from previous learning—such as the US omission during extinction trials—may lead to a new latent cause being inferred, for instance a latent cause that predicts the CS+ but not the US (e.g., an "extinction" cause).

Here, the latent cause model proposed by Gershman and colleagues (Gershman & Hartley, 2015; Gershman & Niv, 2012; see also https://github.com/sjgershm/LCM) was considered. This model assumes that individuals compute on each trial the posterior probability that a given latent cause c generated the observed pattern of CS and US (i.e., presence or absence of the CS+, CS-, and US) using Bayes' rule (for more technical detail, see the supplementary information in Gershman & Hartley, 2015):

$$P(cause = c|stimuli) \propto P(stimuli|cause = c) \times P(cause = c)$$

The inferred probability of cause c being active on a given trial given the observed stimulus configuration ($P(cause = c|stimuli)$) is proportional to the product of the likelihood of that cause ($P(stimuli|cause = c)$) and the prior ($P(cause = c)$). The likelihood of cause c expresses the consistency between the current stimuli and the predicted stimulus configuration associated with this cause. The prior expresses an individual's preference for simpler or more complex causal structures (Norbury et al., 2022). It biases the model to assign new trials to a given cause proportionally to the number of previous trials assigned to that cause, and to a new cause with a probability proportional to the free parameter α. Specifically, α is the concentration parameter of a Chinese restaurant process that models the distribution over latent causes (Gershman & Hartley, 2015; Gershman & Niv, 2012). Lower values of α lead individuals to assign trials to a small number of latent causes, whereas higher values lead individuals to assign trials to a new latent cause. For the fitting procedure, particle filtering with 1000 particles was used to approximate Bayesian inference (Gershman & Hartley, 2015). The concentration parameter α was bounded within the range [0, 10] and the maximum number of latent causes was fixed to 10.

2.5.2. Model and parameter fitting

The computational models were fitted to the trial-by-trial normalized (i.e., scaled and square-root-transformed) SCR to the CSs+ and CSs-. The trial-by-trial normalized SCR was mapped onto the trial-by-trial timeseries of expected values V_t to fit the Rescorla-Wagner model variants. The hybrid model variants were fitted in a threefold manner by mapping the trial-by-trial normalized SCR onto (a) the trial-by-trial timeseries of values V_t, (b) the trial-by-trial timeseries of associabilities κ_t, or (c) the combination of both values and associabilities (Homan et al., 2019; Li et al., 2011; Tzovara et al., 2018; Zhang et al., 2016). As participants were expecting to receive electric stimulations at the outset of the threat conditioning procedure (because of the electric stimulation calibration and the instructions), each initial CS expected value V_0 was set to 0.5 in these models. The trial-by-trial normalized SCR was mapped onto the trial-by-trial US prediction timeseries to fit the latent cause model.

The free parameters were estimated using maximum a posteriori estimation. This estimation procedure consisted in finding the set of parameters maximizing the likelihood of each participant's trial-by-trial normalized SCRs to the CS given the model, constrained by a regularizing prior (Gershman, 2016; Niv et al., 2012; see supplementary materials). The mfit toolbox (Gershman, 2016; https://github.com/sjgershm/mfit) in MATLAB R2021b was used to estimate the free parameters. A separate set of free parameters was estimated for each participant across each stimulus category (Boll et al., 2013). This enabled a comparison of the parameter estimates that best fitted the normalized SCR data between the threat-relevant, positive-relevant, and neutral stimuli.

2.5.3. Model comparison

Model comparison was conducted using random-effects Bayesian model selection (Rigoux et al., 2014) to identify the model most likely to have generated the observed SCR data. This procedure assumes that each participant is drawn from a single population distribution over models, which is estimated from the sample of model evidence values for each model (Gershman, 2016). The model evidence values were calculated by means of the Akaike information criterion (AIC). The AIC was used because it considers both model complexity and goodness of fit, balancing parsimony and predictive accuracy. The protected exceedance probability was computed as a metric to compare the models. It corresponds to the probability that a given model is more frequent in the population than all the other models under consideration while accounting for the possibility that some differences in model evidence may be due to chance. Other model comparison metrics are reported in the supplementary materials.

A model and parameter recovery analysis (Correa et al., 2018; Palminteri et al., 2017; Wilson & Collins, 2019) was additionally performed. This analysis aimed to ensure that the models included in the model space were identifiable and that their parameters can be reliably recovered (see supplementary materials).

2.6. Statistical analyses

All statistical analyses were performed with R (version 4.4.2; R Core Team, 2024; RRID:SCR_001905) and RStudio (version 2024.12.1+563; Posit team, 2024; RRID:SCR_000432). The packages *tidyverse* (Wickham et al., 2019) and *ggplot2* (Wickham, 2016) were used for data wrangling and visualization, respectively.

2.6.1. Skin conductance response

Following standard practice in the human conditioning literature (Lonsdorf et al., 2017), the SCR data was analyzed separately for each conditioning phase. Linear mixed-effects models (LMMs) with the *lme4* (Bates et al., 2015) and *lmerTest* (Kuznetsova et al., 2017) packages were used to analyze each conditioning phase. For the habituation and extinction phases, the within-participants factors CS category (threat--relevant vs. positive-relevant vs. neutral), CS type (CS+ vs. CS-), and their interaction were entered as fixed effects. For acquisition, the acquisition trials were split into an early (i.e., first three presentations of each CS following the first pairing between the CS+ from the stimulus category and the US) and a late (i.e., the subsequent three presentations of each CS) phase to specifically examine the effects of faster acquisition. The within-participants factor subphase and the three-way interaction were additionally entered as fixed effects in the acquisition LMM. The random-effects structure was modeled by first considering a maximal structure (Barr et al., 2013) including random intercepts for participants and by-participant random slopes for each fixed effect. When the maximal model led to singularity, indicating overfitting (Bates et al., 2018), the random-effects structure was sequentially simplified until there was no singular fit by (1) removing the correlation among by-participant random effects (zero-correlation parameter), (2) modeling random-intercepts at each level of the within-participants factors, (3) removing the zero-estimated random effects detected using a principal component analysis of the random-effects covariance matrix estimates, and (4) only modeling random intercepts for participants. The final model was then compared with the maximal model to ensure that the same qualitative results were observed between both models (see supplementary materials). The 'bobyqa' optimizer was used to fit the LMMs with a maximal number of model iterations set to 1 million. The degrees of freedom and the p-values were computed using the Kenward-Roger method. Partial omega squared (ω_p^2) and their 90% confidence interval (CI) are reported as estimates of effect sizes and were calculated using the *effectsize* package (Ben-Shachar et al., 2020).

Planned contrast analyses were conducted with the *emmeans* package (Lenth et al., 2024) to test the a priori hypotheses that the difference in SCR between the CS+ and the CS- is greater during early acquisition—reflecting faster acquisition—and during extinction—reflecting enhanced resistance to extinction—for both threat- and positive-relevant stimuli compared to neutral stimuli. Following this main contrast, three pairwise contrasts were performed to compare (a) threat-relevant versus neutral stimuli, (b) positive-relevant versus neutral stimuli, and (c) threat-relevant versus positive-relevant stimuli. Because these contrasts were nonorthogonal, a Holm-Bonferroni sequential procedure (Holm, 1979) was applied to correct for multiple testing. One-sided testing was performed to test the theory-driven directional predictions (main contrast and contrasts a and b) and two-sided testing was used when there was no directional prediction (contrast c; see Stussi et al., 2018; Stussi, Pourtois, et al., 2021). Cohen's d for LMMs (d_{LMM}; Westfall et al., 2014) and their 95% CI are reported as effect size estimates for the planned contrast analyses. For each contrast, a Bayes factor (BF_{10}) was additionally computed using the *brms* (Bürkner, 2017) and *cmdstanr* (Gabry et al., 2024) packages. The BF_{10} quantifies the likelihood of the data under the alternative hypothesis relative to the likelihood of the data under the null hypothesis. The Bayesian models were estimated using Markov chain Monte Carlo (MCMC) sampling with 4 chains of 10000 iterations and a warmup of 2500 iterations. Prior parameters were set as Cauchy(0, 0.5) distributions.

Additionally, a robustness analysis was conducted by calculating the CR as the SCR to the CS+ minus the SCR to the CS- from the same stimulus category (Olsson et al., 2005; Stussi et al., 2015, 2018, 2019; Stussi, Pourtois, et al., 2021). The results of this analysis showed qualitatively and quantitatively highly consistent results with the LMM approach (see supplementary materials).

2.6.2. *Computational modeling parameters*

Based on the results of the model comparison procedure (see section 3.2.1), the estimated excitatory and inhibitory learning rates extracted from the dual-learning rate Rescorla-Wagner model were analyzed with separate one-way repeated-measures analyses of variance (ANOVAs) with CS category (threat-relevant vs. positive-relevant vs. neutral) as a within-participants factor using the *afex* package (Singmann et al., 2024). Because the residuals were not normally distributed, robust repeated-measures ANOVAs were additionally conducted using the *WRS2* package (Mair & Wilcox, 2020). A priori hypotheses were tested by comparing the excitatory and inhibitory learning rates associated with threat- and positive-relevant stimuli versus neutral stimuli using the same planned contrast analysis as described for the SCR data. ω_p^2 and d_{av} and their 90% or 95% CI are reported as effect size estimates (Lakens, 2013) for the ANOVAs and planned contrasts, respectively. A BF_{10} was computed for each contrast using the same specifications as for the skin conductance response data.

3. Results

3.1. *Skin conductance response*

The trial-by-trial average SCR magnitudes to threat-relevant, positive-relevant, and neutral stimuli across the habituation, acquisition, and extinction phases are depicted in Fig. 1.

3.1.1. *Habituation*

Analysis of the habituation phase (Table 1) indicated that there were no preexisting differences in SCR to the different stimulus categories and no interaction effect with the CS types was observed. However, there was a statistically significant main effect of CS type. SCRs were higher in response to the CSs+ ($M = 0.36$, $SD = 0.31$) than the CSs- ($M = 0.32$, $SD = 0.29$) during habituation (see Fig. S1 in the supplementary materials). Follow-up analyses revealed that this difference was only observed for the first presentation of the CSs ($M_{CS+} = 0.45$ vs. $M_{CS-} = 0.37$; $p < .001$) but not for the second one ($M_{CS+} = 0.27$ vs. $M_{CS-} = 0.27$; $p = .977$). These

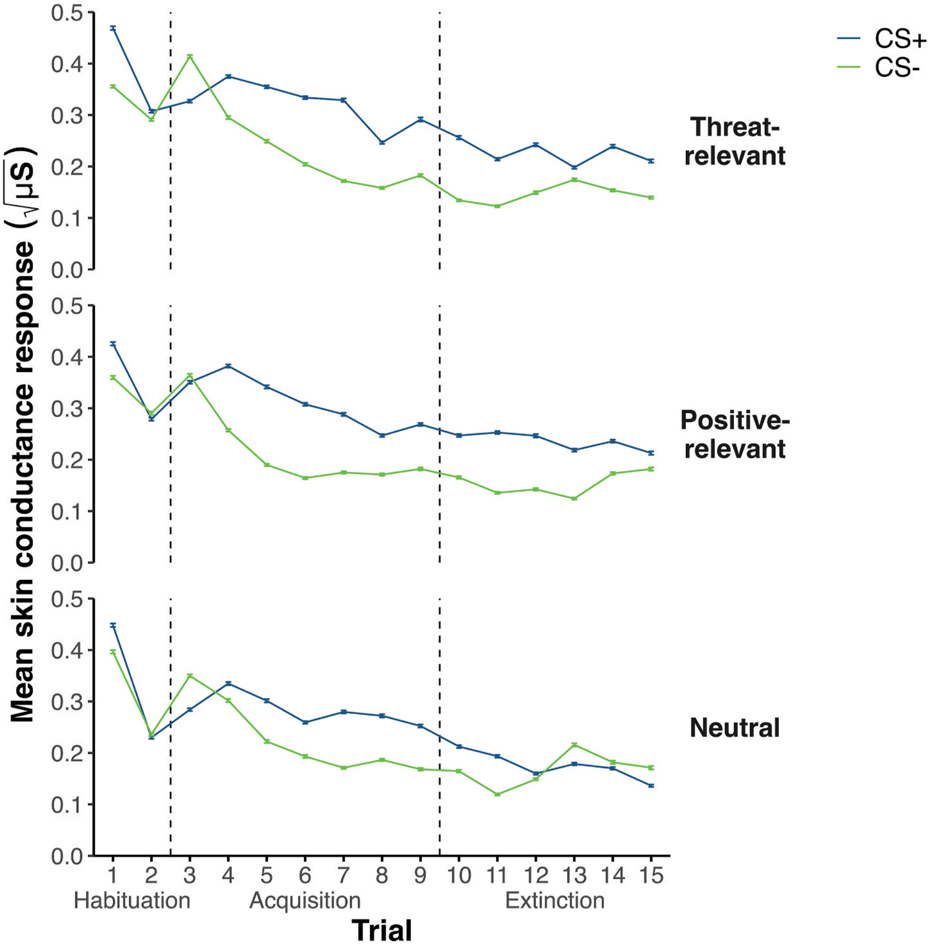

Fig. 1. Trial-by-trial mean skin conductance response to the conditioned stimuli.
Note. $N = 247$. Mean scaled and square-root-transformed skin conductance response as a function of the conditioned stimulus type (CS+ vs. CS-) and category (threat-relevant vs. positive-relevant vs. neutral) across trials. Error bars indicate 95% confidence intervals. CS+ = reinforced conditioned stimulus, CS- = unreinforced conditioned stimulus.

Table 1
Fixed effects from the linear mixed-effects model on the skin conductance response data during habituation.

Fixed effect	Sum of squares	MSE	df	*F*-value	*p*-value	ω_p^2	90 % CI
CS category	0.40	0.20	2, 2705.20	1.76	.173	.0006	[.000, .002]
CS type	1.10	1.10	1, 2705.20	9.69	.002	.003	[.001, .008]
CS category × CS type	0.26	0.13	2, 2705.20	1.13	.324	.0001	[.000, .001]
Model equation							
$SCR_{normalized} \sim CScategory \times CStype + (1 \mid participant_{id})$							

Note. MSE = mean squared error, df = degrees of freedom, ω_p^2 = partial omega squared, CI = confidence interval.

results suggest that the CSs+ might have been more likely to elicit an orienting response than the CSs- because their first presentation occurred earlier on average (M = 3.16 trials) than the first CS- presentation (M = 3.97 trials) during habituation.

3.1.2. Acquisition

Analysis of acquisition (Table 2) revealed a main effect of CS type. As expected, SCRs were higher to the CSs+ (M = 0.30, SD = 0.22) than to the CSs- (M = 0.20, SD = 0.19), indicating successful differential conditioning (see Fig. 2). The main effect of CS category reached statistical significance. Threat-relevant stimuli (M = 0.27, SD = 0.21) evoked slightly higher SCRs than positive-relevant (M = 0.25, SD = 0.22) and neutral (M = 0.25, SD = 0.21) stimuli. There was also a main effect of subphase, with greater SCRs during early (M = 0.28, SD = 0.27) than late (M = 0.23, SD = 0.24) acquisition trials. Additionally, a statistically significant interaction effect between CS type and CS category emerged. These effects were however qualified by the three-way interaction between CS type, CS category, and subphase.

The CR to both threat-relevant and positive-relevant stimuli was acquired faster than to neutral stimuli, as indicated by a greater CS+/CS- difference in SCR during early acquisition, $t(1474) = 3.18$, $p < .001$ (one-tailed), $d_{LMM} = 0.358$ (Table 3; Fig. 3). Direct comparisons revealed a faster CR acquisition to threat-relevant compared to neutral stimuli, t (1476) = 1.98, $p = .024$ (one-tailed), $d_{LMM} = 0.128$, although evidence for this difference was inconclusive ($BF_{10} = 0.837$). Positive-relevant stimuli led to a faster CR acquisition to relative to neutral stimuli, t (1471) = 3.54, $p < .001$ (one-tailed), $d_{LMM} = 0.230$. By contrast, no statistically significant difference was observed between threat-relevant and positive-relevant stimuli, $t(1474) = -1.56$, $p = .119$, $d_{LMM} = -0.101$, with moderate evidence for an absence of such difference ($BF_{01} = 1/BF_{10} = 5.128$).

3.1.3. Extinction

Analysis of the extinction phase (Table 4) showed a statistically significant main effect of CS type, reflecting higher SCRs to the CSs+ (M = 0.21, SD = 0.22) than the CSs- (M = 0.16, SD = 0.17) during the extinction phase. In addition, there was a main effect of CS category. Threat-relevant (M = 0.19, SD = 0.20) and positive-relevant (M = 0.20, SD = 0.20) stimuli were associated with slightly higher SCRs than neutral stimuli (M = 0.17, SD = 0.19). These main effects were qualified by their interaction, indicating that the CS categories differentially influenced the CS+/CS- differentiation during extinction.

The CR to both threat-relevant and positive-relevant stimuli was more persistent than to neutral stimuli, as reflected by a higher CS+/CS- difference in SCR, $t(8333) = 5.38$, $p < .001$, $d_{LMM} = 0.431$ (Table 5; Fig. 4). More focused comparisons showed a greater persistence of the CR to threat-relevant stimuli, $t(8333) = 4.64$, $p < .001$ (one-tailed), $d_{LMM} = 0.215$, and positive-relevant stimuli, $t(8333) = 4.68$, $p < .001$ (one-tailed), $d_{LMM} = 0.217$, versus neutral stimuli. Conversely, no statistical difference in the CR resistance to extinction was found between threat-relevant and positive-relevant stimuli, $t(8333) = -0.05$, $p = .964$, $d_{LMM} = -0.002$, with strong evidence in favor of an absence of such difference ($BF_{01} = 19.608$).

3.2. Computational modeling

3.2.1. Model comparison

Bayesian model selection using the AIC as model evidence identified the dual-learning-rate Rescorla-Wagner model as the most likely to have generated the observed trial-by-trial normalized SCR data (Fig. 4) across the threat-relevant, positive-relevant, and neutral CS categories. This model was associated with a higher protected exceedance probability, indicating that this model was more prevalent in the sample than the other models with a high probability. As a result, the estimated excitatory and inhibitory learning rates from this model were extracted and compared between the threat-relevant, positive-relevant, and neutral stimuli.

3.2.2. Learning parameters

Excitatory learning rates. Analysis of the estimated excitatory learning rates (Table 6) showed no statistically significant main effect of CS category. Similarly, the robust repeated-measures ANOVA did not reveal a statistically significant main effect of CS category.

The planned contrast analysis (Table 7) did not show any statistically significant differences in excitatory learning rates between threat-relevant, positive-relevant, and neutral stimuli after correcting for

Table 2
Fixed effects from the linear mixed-effects model on the skin conductance response data during acquisition.

Fixed effect	Sum of squares	MSE	df	*F*-value	*p*-value	ω_p^2	90 % CI
Subphase	3.40	3.40	1, 245.95	36.52	<.001	.125	[.068, .192]
CS category	0.74	0.37	2, 491.72	3.95	.020	.012	[.000, .030]
CS type	14.41	14.41	1, 245.93	154.76	<.001	.383	[.308, .451]
Subphase × CS category	0.26	0.13	2, 1474.91	1.41	.244	.001	[.000, .003]
Subphase × CS type	0.00	0.00	1, 245.87	0.004	.950	.000	[.000, .000]
CS category × CS type	0.68	0.34	2, 1474.05	3.67	.026	.004	[.000, .010]
Subphase × CS category × CS type	0.65	0.32	2, 1474.03	3.48	.031	.003	[.000, .009]
Model equation							
$SCR_{normalized} \sim subphase \times CScategory \times CStype + (1 \mid participant_{id}) + (1 \mid participant_{id} : subphase) + (1 \mid participant_{id} : CScategory) + (1 \mid participant_{id} : CStype) + (1 \mid participant_{id} : subphase : CStype) + (1 \mid participant_{id} : subphase : CScategory : CStype)$							

Note. MSE = mean squared error, df = degrees of freedom, ω_p^2 = partial omega squared, CI = confidence interval.

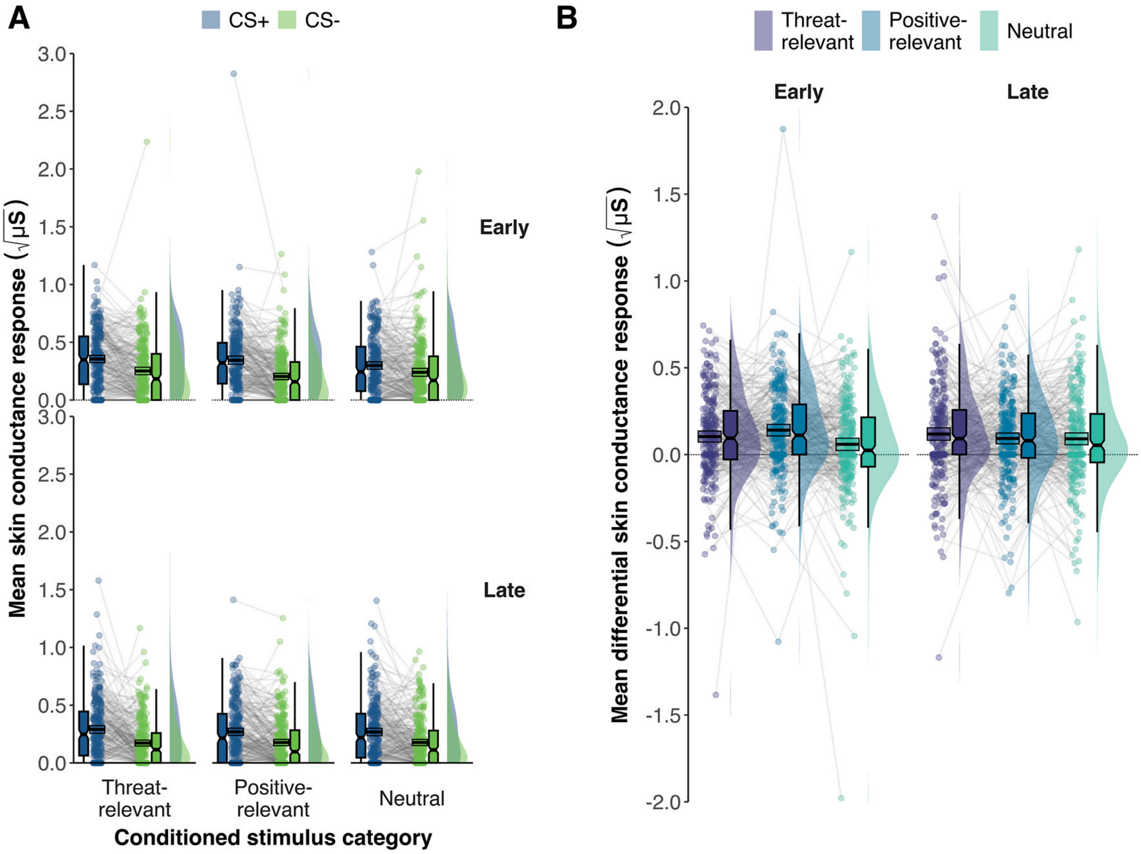

Fig. 2. Mean skin conductance response to the conditioned stimuli during acquisition.
Note. $N = 247$. (A) Mean scaled and square-root-transformed skin conductance response as a function of the acquisition subphase (early vs. late), conditioned stimulus type (CS+ vs. CS-), and conditioned stimulus category (threat-relevant vs. positive-relevant vs. neutral). (B) Mean differential scaled and square-root-transformed skin conductance response (CS+ minus CS-) as a function of the conditioned stimulus category during early and late acquisition. The dots and distributions represent individual participants' data, the crossbars represent the means and their 95% confidence intervals, and the boxplots represent the medians, the interquartile ranges, and ±1.5 interquartile ranges.

Table 3
Results from the planned contrast analysis on the skin conductance response data during early acquisition.

Contrast	Estimate	SE	df	*t*-value	*p*-value	d_{LMM}	95 % CI	BF_{10}
Threat + Pos vs. Neu	0.13	0.04	1474	3.18	<.001[a]	0.358	[0.137, 0.579]	11.954
Threat vs. Neu	0.05	0.02	1476	1.98	.024[a]	0.128	[0.001, 0.256]	0.837
Pos vs. Neu	0.08	0.02	1471	3.54	<.001[a]	0.230	[0.102, 0.357]	51.097
Threat vs. Pos	−0.04	0.02	1474	−1.56	.119	−0.101	[-0.229, 0.026]	0.195

Note. Threat = threat-relevant stimuli, Pos = positive-relevant stimuli, Neu = neutral stimuli, SE = standard error, df = degrees of freedom, d_{LMM} = Cohen's *d* for linear mixed-effects models, CI = confidence interval, BF_{10} = Bayes factor comparing the alternative hypothesis (H1) to the null hypothesis (H0).
[a] Indicates one-tailed testing.

multiple testing (α = .0125; all *ps* > .019).

Inhibitory learning rates. The CS categories differentially influenced the estimated inhibitory learning rates (Table 6). This main effect was likewise found with the robust repeated-measures ANOVA.

Threat-relevant and positive-relevant stimuli were associated with lower inhibitory learning rates compared to neutral stimuli, $t(243) = 2.93$, $p = .002$ (one-tailed), $d_{av} = 0.240$ (Table 8; Fig. 5). Focused comparisons indicated that the estimated inhibitory learning rates for threat-relevant stimuli were lower than for neutral stimuli, $t(243) = 2.96$, $p = .002$ (one-tailed), $d_{av} = 0.250$. These estimates were also lower for positive-relevant relative to neutral stimuli, $t(243) = 2.11$, $p = .018$ (one-tailed), $d_{av} = 0.176$, but evidence for this difference was anecdotal ($BF_{10} = 1.559$). No statistically significant difference in estimated inhibitory learning rates was observed between threat-relevant and positive-relevant stimuli, $t(243) = 0.73$, $p = .467$, $d_{av} = 0.063$, with moderate evidence for the absence of such difference ($BF_{01} = 8.997$).

4. Discussion

By combining and reanalyzing data from four experiments on Pavlovian threat conditioning (Stussi et al., 2018; Stussi, Pourtois, et al., 2021), this study aimed to provide a confirmatory analysis of prior findings supporting the appraisal-based hypothesis that Pavlovian learning biases are driven by an affective relevance mechanism that is not specific to threat, and further characterize their computational signatures. Overall, results indicate that both threat- and positive-relevant stimuli were more readily and persistently associated with an aversive outcome during Pavlovian threat conditioning than neutral stimuli. While no evidence was found that threat acquisition biases were linked to a higher excitatory learning rate, both threat- and positive-relevant stimuli exhibited a lower inhibitory learning rate compared to neutral stimuli, contributing to the persistence of the conditioned response during extinction. These results confirm and support the robustness of

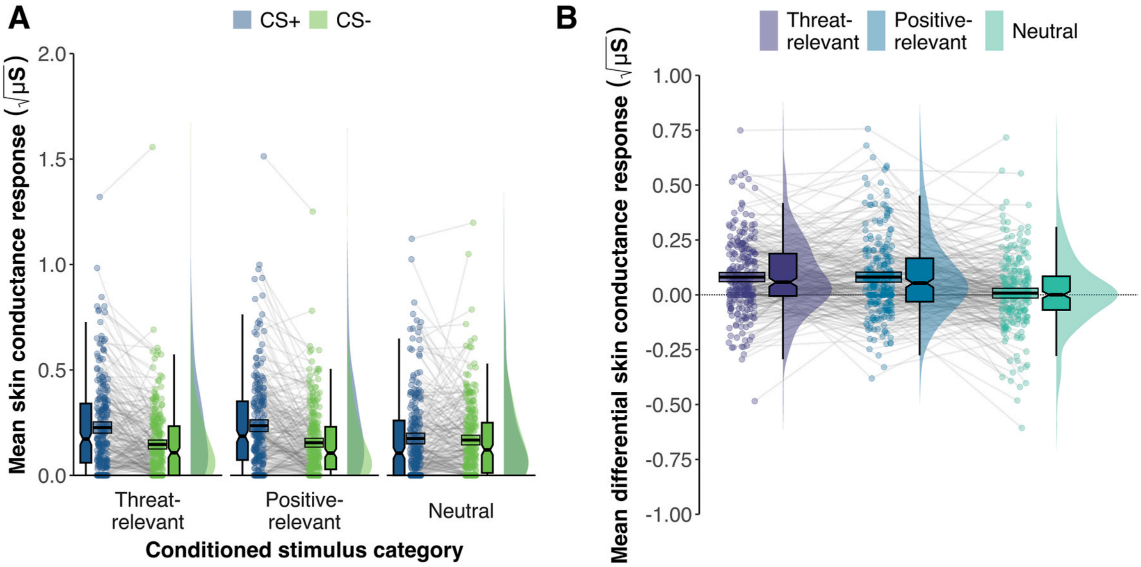

Fig. 3. Mean skin conductance response to the conditioned stimuli during extinction.
Note. N = 247. (A) Mean scaled and square-root-transformed skin conductance response as a function of the conditioned stimulus type (CS+ vs. CS-) and category (threat-relevant vs. positive-relevant vs. neutral). (B) Mean differential scaled and square-root-transformed skin conductance response (CS+ minus CS-) as a function of the conditioned stimulus category. The dots and distributions represent individual participants' data, the crossbars represent the means and their 95% confidence intervals, and the boxplots represent the medians, the interquartile ranges, and ±1.5 interquartile ranges.

Table 4
Fixed effects from the linear mixed-effects model on the skin conductance response data during extinction.

Fixed effect	Sum of squares	MSE	df	*F*-value	*p*-value	ω_p^2	90 % CI
CS category	0.86	0.43	2, 8333.60	4.73	.009	.001	[.0001, .002]
CS type	4.71	4.71	1, 245.90	51.70	<.001	.170	[.105, .240]
CS category × CS type	2.64	1.32	2, 8333.20	14.48	<.001	.003	[.001, .005]
Model equation $SCR_{normalized} \sim CScategory \times CStype + (1\|participant_{id}) + (1\|participant_{id} : CStype)$							

Note. MSE = mean squared error, df = degrees of freedom, ω_p^2 = partial omega squared, CI = confidence interval.

Table 5
Results from the planned contrast analysis on the skin conductance response data during extinction.

Contrast	Estimate	SE	df	*t*-value	*p*-value	d_{LMM}	95 % CI	BF_{10}
Threat + Pos vs. Neu	0.15	0.03	8333	5.38	<.001[a]	0.431	[0.274, 0.588]	6.72×10^{16}
Threat vs. Neu	0.07	0.02	8333	4.64	<.001[a]	0.215	[0.124, 0.305]	2.71×10^{7}
Pos vs. Neu	0.07	0.02	8333	4.68	<.001[a]	0.217	[0.126, 0.307]	1.67×10^{7}
Threat vs. Pos	−0.00	0.02	8333	−0.05	.964	−0.002	[-0.093, 0.089]	0.051

Note. Threat = threat-relevant stimuli, Pos = positive-relevant stimuli, Neu = neutral stimuli, SE = standard error, df = degrees of freedom, d_{LMM} = Cohen's *d* for linear mixed-effects models, CI = confidence interval, BF_{10} = Bayes factor comparing the alternative hypothesis (H1) to the null hypothesis (H0).
[a] Indicates one-tailed testing.

the original findings reported by Stussi et al. (Stussi et al., 2018; Stussi, Pourtois, et al., 2021).

The pattern of results for threat-relevant stimuli are consistent with previous research supporting the preparedness and fear module theories (Ho & Lipp, 2014; Olsson et al., 2005; Öhman & Dimberg, 1978; Öhman et al., 1976; Öhman & Mineka, 2001). However, the similar effects observed for positive-relevant stimuli question the notion that enhanced Pavlovian threat conditioning is selective to threat-relevant stimuli. Instead, these results suggest that learning biases during Pavlovian threat acquisition and extinction also occur in response to positive stimuli with heightened affective value. This challenges the preparedness and fear module theories, while lending support for the affective relevance model derived from appraisal theories of emotion. According to this model, Pavlovian learning biases arise in response to stimuli appraised as affectively relevant to the individual, regardless of their valence and inherent threat value (Stussi et al., 2015, 2018, 2019; Stussi, Pourtois, et al., 2021). These findings contribute to informing the reproducibility and robustness of affective relevance's effects on Pavlovian threat conditioning, thereby further elucidating the affective mechanisms that influence Pavlovian threat acquisition and extinction.

Given that defensive responses differ for actual dangers compared to standard laboratory-based threat conditioning paradigms (Mobbs & Kim, 2015), it could be argued that the lack of differential conditioning effects between threat- and positive-relevant stimuli arose because viewing threat-relevant stimuli in a laboratory setting poses no actual threat, unlike real-life encounters. While the affective relevance model predicts that real-life threats are likely to produce stronger conditioning effects due to their higher relevance to survival-related concerns, it also

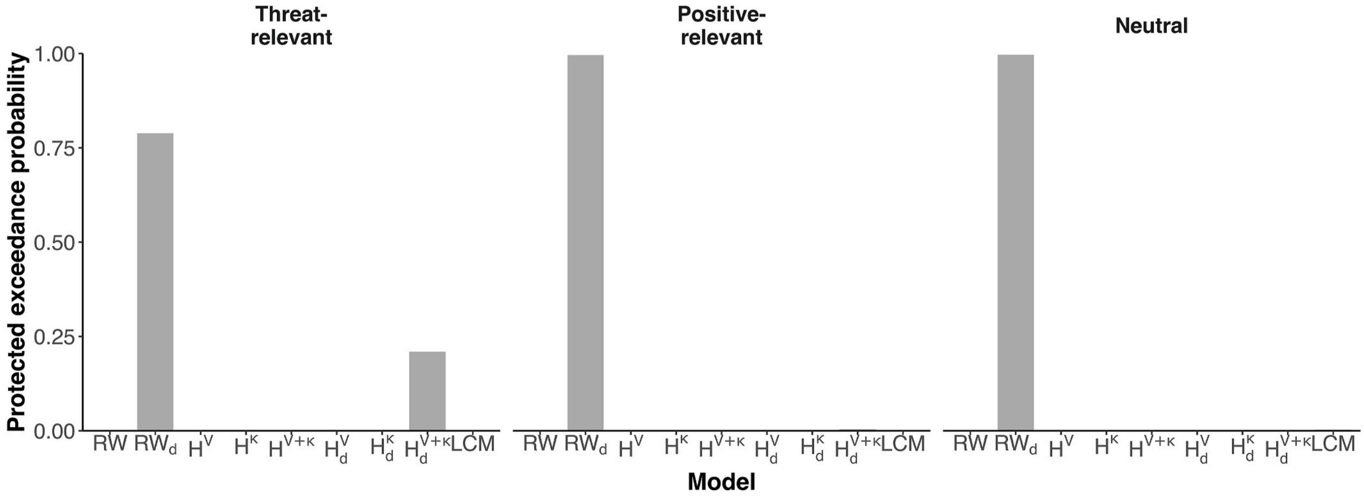

Fig. 4. Results from the Bayesian model selection for each conditioned stimulus category.
Note. The Bayesian model selection procedure used the Akaike information criterion as model evidence and protected exceedance probability to identify the most probable model to have generated the skin conductance response data across each stimulus category. RW = Rescorla-Wagner model, d = dual learning rates, H = hybrid model, V = expected value, κ = associability, LCM = latent cause model.

Table 6
Results from the repeated-measures analyses of variance on the estimated learning rates.

	Factor	Estimation	Sum of squares	MSE	df	*F*-value	*p*-value	ω_p^2	90 % CI
η^+	CS category	LS	0.29	0.07	2, 486	2.12	.122	.002	[.000, .012]
		Robust	–	–	1.98, 291.18	0.95	.385	–	–
η^-	CS category	LS	0.68	0.08	2, 486	4.48	.012	.008	[.000, .024]
		Robust	–	–	1.98, 291.26	4.03	.019	–	–

Note. η^+ = excitatory learning rate, η^- = inhibitory learning rate, LS = least-squares estimation, MSE = mean squared error, df = degrees of freedom, ω_p^2 = partial omega squared, CI = confidence interval.

Table 7
Results from the planned contrast analysis on the estimated excitatory learning rates.

Contrast	Estimate	SE	df	*t*-value	*p*-value	d_{av}	95 % CI	BF_{10}
Threat + Pos vs. Neu	0.08	0.04	243	1.91	.029[a]	0.144	[-0.004, 0.293]	0.372
Threat vs. Neu	0.03	0.02	243	1.17	.122[a]	0.093	[-0.063, 0.250]	0.245
Pos vs. Neu	0.05	0.02	243	2.07	.020[a]	0.162	[0.008, 0.316]	1.020
Threat vs. Pos	−0.02	0.02	243	−0.89	.376	−0.070	[-0.225, 0.085]	0.079

Note. Threat = threat-relevant stimuli, Pos = positive-relevant stimuli, Neu = neutral stimuli, SE = standard error, df = degrees of freedom, d_{av} = averaged Cohen's d, CI = confidence interval, BF_{10} = Bayes factor comparing the alternative hypothesis (H1) to the null hypothesis (H0).

[a] Indicates one-tailed testing.

Table 8
Results from the planned contrast analysis on the estimated inhibitory learning rates.

Contrast	Estimate	SE	df	*t*-value	*p*-value	d_{av}	95 % CI	BF_{10}
Threat + Pos vs. Neu	0.12	0.04	243	2.93	.002[a]	0.240	[0.078, 0.402]	6.792
Threat vs. Neu	0.07	0.02	243	2.96	.002[a]	0.250	[0.083, 0.417]	10.490
Pos vs. Neu	0.05	0.03	243	2.11	.018[a]	0.176	[0.012, 0.340]	1.559
Threat vs. Pos	0.02	0.03	243	0.73	.467	0.063	[-0.107, 0.234]	0.111

Note. Threat = threat-relevant stimuli, Pos = positive-relevant stimuli, Neu = neutral stimuli, SE = standard error, df = degrees of freedom, d_{av} = averaged Cohen's d, CI = confidence interval, BF_{10} = Bayes factor comparing the alternative hypothesis (H1) to the null hypothesis (H0).

[a] Indicates one-tailed testing.

predicts similar effects for more ecological positive-relevant stimuli. Relatedly, the current findings could be seen as reflecting the involvement of two different mechanisms providing convergent outputs at the skin conductance and algorithmic levels: a threat-specific mechanism preferentially processing threat-relevant stimuli, and another dedicated to positive or nonthreatening affectively relevant stimuli. Accordingly, future research could test the effects of affective relevance on Pavlovian threat conditioning using more ecologically valid paradigms (e.g., virtual reality; Kredlow et al., 2022) that mimic the conditions under which threats and rewards typically occur in the environment within a neurocomputational framework (Mobbs et al., 2021). This would contribute to establishing whether biases in Pavlovian threat acquisition and extinction to threat- and positive-relevant stimuli rely on shared or functionally distinct mechanisms.

Crucially, computational analyses identified the Rescorla-Wagner model with dual learning rates for excitatory and inhibitory learning

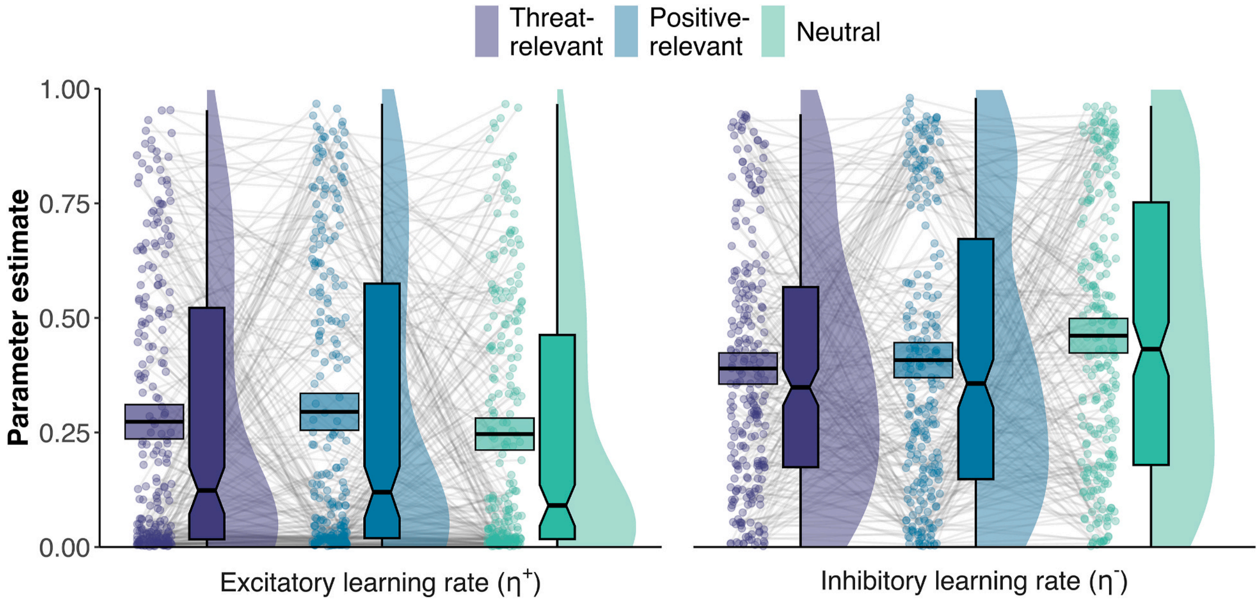

Fig. 5. Estimated excitatory and inhibitory learning rates.
Note. $N = 244$. Estimated learning rates derived from the dual-learning-rate Rescorla-Wagner model as a function of the conditioned stimulus category (threat-relevant vs. positive-relevant vs. neutral). The dots and distributions represent individual participants' data, the crossbars represent the means and their 95% confidence intervals, and the boxplots represent the medians, the interquartile ranges, and ±1.5 interquartile ranges.

as more likely to have generated the observed skin conductance data. This model was favored over alternative models that do not distinguish between excitatory and inhibitory learning, integrate a dynamic learning rate, or separate threat acquisition and extinction through a latent cause inference process rather than learning rates. These results somewhat differ from prior studies that identified the hybrid model implementing an associability-driven dynamic learning rate as the best fit for SCR data in human differential threat conditioning paradigms (Homan et al., 2019; Li et al., 2011; Oyarzun et al., 2022; Zhang et al., 2016). In these studies, conditioned SCR better mapped onto stimulus associability (Li et al., 2011; Zhang et al., 2016; see also Tzovara et al., 2018) or the combination of stimulus associability and value (Homan et al., 2019; Oyarzun et al., 2022) than stimulus value alone. These differences may stem from variations in experimental design (e.g., extinction vs. reversal learning), the nature and number of stimuli used, or the model space tested as dual-learning-rates variants of the Rescorla-Wagner and hybrid models were often not included. Investigating these factors systematically could help reconcile these discrepancies and identify the conditions under which conditioned SCR is better explained by a dynamic learning rate (Homan et al., 2019; Li et al., 2011; Zhang et al., 2016), dual static learning rates (Starita et al., 2023; Stussi et al., 2018; Stussi, Pourtois, et al., 2021), a combination thereof (Oyarzun et al., 2022), or alternative models (Gershman & Hartley, 2015; Tzovara et al., 2018).

Regarding the estimated learning parameters, threat- and positive-relevant stimuli were both associated with a lower inhibitory learning rate than neutral stimuli, confirming preliminary evidence reported by Stussi and colleagues (Stussi et al., 2018; Stussi, Pourtois, et al., 2021). This lower inhibitory learning rate reduced the integration of negative prediction errors, thereby altering extinction learning and enhancing the persistence of the conditioned response. Bayesian analyses on direct pairwise comparisons suggested that the evidence for the difference in inhibitory learning rates between threat-relevant and neutral stimuli was strong, whereas the evidence for the difference between positive-relevant and neutral stimuli was only anecdotal. However, there was moderate evidence supporting the absence of a difference between threat- and positive-relevant stimuli. Overall, lower inhibitory learning rates provide a computational mechanism by which affectively relevant stimuli can modulate Pavlovian threat extinction.

Conversely, no evidence was found that faster acquisition to threat- and positive-relevant stimuli compared to neutral stimuli was associated with increased excitatory learning rates. This contrasts with previous work showing heightened learning rates for threat-relevant stimuli during aversive reversal learning (Atlas & Phelps, 2018). A potential explanation for this discrepancy is that excitatory learning rates were estimated across the whole acquisition phase, whereas differences between affectively relevant and neutral stimuli specifically occurred during early acquisition. As such, excitatory learning rates might capture a mixture between faster and stronger threat acquisition rather than pure effects of faster acquisition per se. Moreover, habituation effects in skin conductance response—which are caused by repeated exposure to the same stimuli (Society for Psychophysiological Research Ad Hoc Committee on Electrodermal Measures, 2012)—may have masked differences in excitatory learning rates between stimulus categories by influencing the accuracy of their estimation. Even though these explanations are speculative and require caution, further research using less habituation-sensitive psychophysiological measures, such as pupil size (Leuchs et al., 2019), is warranted to determine whether Pavlovian threat acquisition biases are associated with higher excitatory learning rates.

The combination of data from four experiments in the present work offers increased precision in quantifying the effects of affectively relevant stimuli on Pavlovian threat conditioning. The estimated effect sizes—along with their confidence intervals—suggest that the biases induced by threat- and positive-relevant stimuli in Pavlovian threat acquisition, extinction, and inhibitory learning rates are small to moderate (d_{LMM} range = 0.13–0.43; d_{av} range = 0.17–0.45; see also the robustness analysis in supplementary materials). These modest effect sizes might account for prior difficulties in replicating the effects of threat-relevance on Pavlovian threat conditioning (McNally, 1987; Åhs et al., 2018), given that earlier studies often used small-to-moderate sample sizes and between-participants designs. The robustness analysis, focusing on aggregated CS+/CS- differentiation (see supplementary materials), further suggested that the effect sizes for faster acquisition (d_{av} range = 0.17–0.40) were generally smaller than those for enhanced resistance to extinction (d_{av} range = 0.40–0.45), particularly for threat-relevant stimuli. This was also reflected by the Bayes factors, which indicated that evidence for the difference between threat-relevant and neutral stimuli during early acquisition was inconclusive. Similarly, differences in excitatory learning rates (d_{av} range = 0.08–0.14) between affectively relevant and neutral stimuli were of smaller sizes than differences in inhibitory learning rates (d_{av} range = 0.17–0.25). This pattern aligns with the broader human conditioning literature, which has found less consistent support for faster threat acquisition to threat-relevant stimuli compared to enhanced resistance to extinction (for reviews, see McNally, 1987; Öhman & Mineka, 2001). These methodological considerations underscore the importance of conducting well-powered studies to reliably assess the influence of different stimulus categories on Pavlovian threat conditioning.

The current study highlights the importance of modeling distinct learning dynamics for threat acquisition and extinction, especially when investigating stimuli with high affective value. This distinction helps elucidate the mechanisms underlying how specific conditioned stimuli can not only facilitate faster acquisition of defensive responses, but also their persistence during extinction—a pattern not fully explained by classical associative learning models (Le Pelley, 2004; Mackintosh, 1975; Pearce & Hall, 1980; Rescorla & Wagner, 1972). Indeed, these classical models predict that conditioned stimuli leading to faster acquisition should, all else being equal, also lead to faster extinction (Siddle & Bond, 1988). The present work shows that differentiating the impact of positive versus negative prediction errors through the modeling of separate learning rates for excitatory and inhibitory learning is a parsimonious and powerful approach to capture learning biases during Pavlovian threat extinction.

A limitation of this approach, however, is that it is limited in its capacity to characterize more complex phenomena beyond threat acquisition and extinction, including spontaneous recovery, reinstatement, or renewal (i.e., return of fear or relapse phenomena; e.g., Dunsmoor et al., 2015; but see Paskewitz et al., 2022). In this context, latent cause models provide a valuable alternative framework for understanding variations in threat acquisition and extinction dynamics, along with return-of-fear phenomena. According to this framework, individuals infer latent causes that underlie the relationship between the conditioned and the unconditioned stimulus (Gershman et al., 2010; Gershman & Niv, 2012). Specifically, some individuals are more likely to cluster acquisition trials (i.e., when the conditioned and unconditioned stimuli are contingently paired) into a common latent cause and to assign extinction trials to a new latent cause (i.e., formation of a new inhibitory association during extinction), while other individuals tend to group all conditioning trials into a single latent cause (i.e., leading to “unlearning” during extinction; Gershman & Hartley, 2015; Norbury et al., 2022). Although the current reanalysis did not identify the latent cause model as the best explanatory model for our data, this class of models can account for why conditioned responses to specific stimuli are rapidly acquired but resist extinction when acquisition and extinction trials are assigned to different causes or contexts, as well as predict (Gershman & Hartley, 2015) or produce (Gershman et al., 2017) the reemergence of defensive responses.

From a translational perspective, the effects of faster threat acquisition, enhanced resistance to extinction, and decreased inhibitory learning rates related to stimuli with high affective value mirror to some extent core hallmarks of anxiety- and stress-related disorders. In fact, these conditions are typically characterized by rapid development of maladaptive defensive responses (e.g., after exposure to a traumatic event), increased threat-related responses to safety signals, and the persistence of these responses even in the absence of actual danger (e.g., Duits et al., 2015; Homan et al., 2019; Kausche et al., 2025; Kindt, 2014; Seligman, 1971), these symptoms being often linked to excessive excitatory learning and impaired inhibitory learning (Lissek & van Meurs, 2015). A recent meta-analysis (Kausche et al., 2025) suggested that patients with anxiety- and stress-related disorders exhibit amplified physiological and behavioral responses to safety cues during acquisition, along with heightened negative affective reactions to threat cues. These patients also show elevated threat expectancy and negative affective reactions to safety cues during extinction, as well as persistent negative affective reactions to threat cues. These alterations were notably observed across diagnostic categories, with some variations. Acquisition effects were more pronounced in patients with anxiety and obsessive-compulsive disorders, while extinction effects were particularly evident in patients with post-traumatic stress disorder. In that sense, the asymmetry between excitatory and inhibitory (or threat and safety) learning may represent a vulnerability factor for the onset and maintenance of anxiety- and stress-related disorders. Modeling dissociable excitatory and inhibitory learning processes and considering the influence of affective factors thereon might thus contribute to better understanding why and how certain individuals develop maladaptive emotional responses and mental health problems in the aftermath of aversive events (Lonsdorf & Merz, 2017; Wise & Dolan, 2020), as well as help predict treatment outcomes and aid in the development of personalized interventions. Despite these promising possibilities, the practical value of this approach remains to be thoroughly and rigorously evaluated, especially regarding the reliability of the computational phenotypes of heightened excitatory and diminished inhibitory learning. Whereas initial evidence supports the reliability of Pavlovian threat conditioning paradigms across relatively short test-retest intervals (Cooper et al., 2023), psychometric properties of computational measures are still highly debated (Karvelis et al., 2023; Schurr et al., 2024; Vrizzi et al., 2025) and require further validation to establish the translational potential of computational phenotyping.

In conclusion, this study suggests that both threat-relevant and positive-relevant stimuli induce learning biases during Pavlovian threat acquisition and extinction, as reflected by faster acquisition and enhanced resistance to extinction of the conditioned response to these stimuli. Pavlovian threat extinction biases to affectively relevant stimuli were characterized by a lower inhibitory learning rate at the algorithmic level, which fostered the persistence of the conditioned response during extinction. These findings indicate that the affective determinants of preferential Pavlovian threat learning are more flexible than previously thought and may hinge upon the stimulus' appraised affective relevance in relation to the individual's goals and well-being (Stussi et al., 2018; Stussi, Pourtois, et al., 2021). They also illustrate the potential of combining emotion theories and computational modeling in providing insights into the understanding of how humans learn and update the affective value of stimuli in their environment (Pool et al., 2023; Schiller et al., 2024; Stussi & Sander, 2024), as well as impairments in these processes contributing to the emergence and persistence of mental health issues (Wuensch et al., 2021).

Data code availability

The data reported in this manuscript and the code used for data analysis are available on GitLab at https://gitlab.unige.ch/Yoann.Stussi/pavext and on the Open Science Framework at https://doi.org/10.17605/OSF.IO/2XB8R. The raw data used to generate the datasets can be found on the Open Science Framework at https://doi.org/10.17605/OSF.IO/2FYHQ and https://doi.org/10.17605/OSF.IO/DK2NP. The experiments and analyses reported in the manuscript were not preregistered.

Declaration of competing interest

The author declares no competing interests.

Acknowledgements

The author would like to thank Prof. Dr. Eva R. Pool, Prof. Dr. Gilles Pourtois, Prof. Dr. Andreas Olsson, and Prof. Dr. David Sander for helpful discussions and comments. This work was partly supported by an

Early Postdoc. Mobility fellowship (P2GEP1_187911) from the Swiss National Science Foundation to YS. YS is supported by an ERC Starting Grant (INFORL-948671) awarded to Prof. Dr. Maël Lebreton.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.brat.2025.104804.

Data availability

I have shared the link to the data and code in the main manuscript under the section "Data and code availability"

References

Abend, R., Burk, D., Ruiz, S. G., Gold, A. L., Napoli, J. L., Britton, J. C., Michalska, K. J., Shechner, T., Winkler, A. M., Leibenluft, E., Pine, D. S., & Averbeck, B. B. (2022). Computational modeling of threat learning reveals links with anxiety and neuroanatomy in humans. *eLife, 11*, Article e66169. https://doi.org/10.7554/eLife.66169

Åhs, F., Rosén, J., Kastrati, G., Fredrikson, M., Agren, T., & Lundström, J. N. (2018). Biological preparedness and resistance to extinction of skin conductance responses conditioned to fear relevant animal pictures: A systematic review. *Neuroscience & Biobehavioral Reviews, 95*, 430–437. https://doi.org/10.1016/j.neubiorev.2018.10.017

Atlas, L. Y., & Phelps, E. A. (2018). Prepared stimuli enhance aversive learning without weakening the impact of verbal instructions. *Learning & Memory, 25*(2), 100–104. https://doi.org/10.1101/lm.046359.117

Bach, D. R., Sporrer, J., Abend, R., Beckers, T., Dunsmoor, J. E., Fullana, M. A., Gamer, M., Gee, D. G., Hamm, A., Hartley, C. A., Herringa, R. J., Jovanovic, T., Kalisch, R., Knight, D. C., Lissek, S., Lonsdorf, T. B., Merz, C. J., Milad, M., Morriss, J., … Schiller, D. (2023). Consensus design of a calibration experiment for human fear conditioning. *Neuroscience & Biobehavioral Reviews, 148*, Article 105146. https://doi.org/10.1016/j.neubiorev.2023.105146

Barr, D. J., Levy, R., Scheepers, C., & Tily, H. J. (2013). Random effects structure for confirmatory hypothesis testing: Keep it maximal. *Journal of Memory and Language, 68*(3), 255–278. https://doi.org/10.1016/j.jml.2012.11.001

Bates, D., Kliegl, R., Vasishth, S., & Baayen, H. (2018). Parsimonious mixed models. https://doi.org/10.48550/arXiv.1506.04967.

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software, 67*, 1–48. https://doi.org/10.18637/jss.v067.i01

Beckers, T., Hermans, D., Lange, I., Luyten, L., Scheveneels, S., & Vervliet, B. (2023). Understanding clinical fear and anxiety through the lens of human fear conditioning. *Nature Reviews Psychology, 2*(4), 233–245. https://doi.org/10.1038/s44159-023-00156-1

Ben-Shachar, M. S., Lüdecke, D., & Makowski, D. (2020). Effectsize: Estimation of effect size indices and standardized parameters. *Journal of Open Source Software, 5*(56), 2815. https://doi.org/10.21105/joss.02815

Betti, S., Badioli, M., Dalbagno, D., Garofalo, S., di Pellegrino, G., & Starita, F. (2024). Topographically selective motor inhibition under threat of pain. *Pain, 165*(12), 2851. https://doi.org/10.1097/j.pain.0000000000003301

Boll, S., Gamer, M., Gluth, S., Finsterbusch, J., & Büchel, C. (2013). Separate amygdala subregions signal Surprise and predictiveness during associative fear learning in humans. *European Journal of Neuroscience, 37*(5), 758–767. https://doi.org/10.1111/ejn.12094

Brainard, D. H. (1997). The psychophysics toolbox. *Spatial Vision, 10*(4), 433–436. https://doi.org/10.1163/156856897X00357

Bürkner, P.-C. (2017). Brms: An R package for Bayesian multilevel models using stan. *Journal of Statistical Software, 80*, 1–28. https://doi.org/10.18637/jss.v080.i01

Cooper, S. E., Dunsmoor, J. E., Koval, K. A., Pino, E. R., & Steinman, S. A. (2023). Test–retest reliability of human threat conditioning and generalization across a 1-to-2-week interval. *Psychophysiology, 60*(6), Article e14242. https://doi.org/10.1111/psyp.14242

Cooper, S. E., Perkins, E. R., Webler, R. D., Dunsmoor, J. E., & Krueger, R. F. (2024). Integrating threat conditioning and the hierarchical taxonomy of psychopathology to advance the study of anxiety-related psychopathology. *Journal of Psychopathology and Clinical Science, 133*(8), 716–732. https://doi.org/10.1037/abn0000945

Correa, C. M. C., Noorman, S., Jiang, J., Palminteri, S., Cohen, M. X., Lebreton, M., & van Gaal, S. (2018). How the level of reward awareness changes the computational and electrophysiological signatures of reinforcement learning. *Journal of Neuroscience, 38* (48), 10338–10348. https://doi.org/10.1523/JNEUROSCI.0457-18.2018

Craske, M. G., Sandman, C. F., & Stein, M. B. (2022). How can neurobiology of fear extinction inform treatment? *Neuroscience & Biobehavioral Reviews, 143*, Article 104923. https://doi.org/10.1016/j.neubiorev.2022.104923

Craske, M. G., Treanor, M., Conway, C. C., Zbozinek, T., & Vervliet, B. (2014). Maximizing exposure therapy: An inhibitory learning approach. *Behaviour Research and Therapy, 58*, 10–23. https://doi.org/10.1016/j.brat.2014.04.006

Duits, P., Cath, D. C., Lissek, S., Hox, J. J., Hamm, A. O., Engelhard, I. M., van den Hout, M. A., & Baas, J. M. P. (2015). Updated meta-analysis of classical fear conditioning in the anxiety disorders. *Depression and Anxiety, 32*(4), 239–253. https://doi.org/10.1002/da.22353

Dunsmoor, J. E., Niv, Y., Daw, N., & Phelps, E. A. (2015). Rethinking extinction. *Neuron, 88*(1), 47–63. https://doi.org/10.1016/j.neuron.2015.09.028

Ehlers, M. R., & Lonsdorf, T. B. (2022). Data sharing in experimental fear and anxiety research: From challenges to a dynamically growing database in 10 simple steps. *Neuroscience & Biobehavioral Reviews, 143*, Article 104958. https://doi.org/10.1016/j.neubiorev.2022.104958

Faul, F., Erdfelder, E., Lang, A.-G., & Buchner, A. (2007). G*Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavior Research Methods, 39*(2), 175–191. https://doi.org/10.3758/BF03193146

Frijda, N. H. (1986). *The emotions*. Cambridge University Press.

Gabry, J., Češnovar, R., Johnson, A., & Bronder, S. (2024). Cmdstanr: R interface to "CmdStan" Version 0.8.1. https://mc-stan.org/cmdstanr/.

Garcia, J., & Koelling, R. A. (1966). Relation of cue to consequence in avoidance learning. *Psychonomic Science, 4*(1), 123–124. https://doi.org/10.3758/BF03342209

Gershman, S. J. (2016). Empirical priors for reinforcement learning models. *Journal of Mathematical Psychology, 71*, 1–6. https://doi.org/10.1016/j.jmp.2016.01.006

Gershman, S. J., Blei, D. M., & Niv, Y. (2010). Context, learning, and extinction. *Psychological Review, 117*(1), 197–209. https://doi.org/10.1037/a0017808

Gershman, S. J., & Hartley, C. A. (2015). Individual differences in learning predict the return of fear. *Learning & Behavior, 43*(3), 243–250. https://doi.org/10.3758/s13420-015-0176-z

Gershman, S. J., Monfils, M.-H., Norman, K. A., & Niv, Y. (2017). The computational nature of memory modification. *eLife, 6*, Article e23763. https://doi.org/10.7554/eLife.23763

Gershman, S. J., & Niv, Y. (2012). Exploring a latent cause theory of classical conditioning. *Learning & Behavior, 40*(3), 255–268. https://doi.org/10.3758/s13420-012-0080-8

Hamm, A. O., Vaitl, D., & Lang, P. J. (1989). Fear conditioning, meaning, and belongingness: A selective association analysis. *Journal of Abnormal Psychology, 98* (4), 395–406. https://doi.org/10.1037/0021-843X.98.4.395

Ho, Y., & Lipp, O. V. (2014). Faster acquisition of conditioned fear to fear-relevant than to nonfear-relevant conditional stimuli. *Psychophysiology, 51*(8), 810–813. https://doi.org/10.1111/psyp.12223

Holm, S. (1979). A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics, 6*(2), 65–70.

Homan, P., Levy, I., Feltham, E., Gordon, C., Hu, J., Li, J., Pietrzak, R. H., Southwick, S., Krystal, J. H., Harpaz-Rotem, I., & Schiller, D. (2019). Neural computations of threat in the aftermath of combat trauma. *Nature Neuroscience, 22*(3), 470–476. https://doi.org/10.1038/s41593-018-0315-x

Karvelis, P., Paulus, M. P., & Diaconescu, A. O. (2023). Individual differences in computational psychiatry: A review of current challenges. *Neuroscience & Biobehavioral Reviews, 148*, Article 105137. https://doi.org/10.1016/j.neubiorev.2023.105137

Kausche, F., Carsten, H., Sobania, K., & Riesel, A. (2025). Fear and safety learning in anxiety- and stress-related disorders: An updated meta-analysis. *Neuroscience & Biobehavioral Reviews, 169*, Article 105983. https://doi.org/10.1016/j.neurobiorev.2024.105983

Kindt, M. (2014). A behavioural neuroscience perspective on the aetiology and treatment of anxiety disorders. *Behaviour Research and Therapy, 62*, 24–36. https://doi.org/10.1016/j.brat.2014.08.012

Kredlow, M. A., De Voogd, L. D., & Phelps, E. A. (2022). A case for translation from the clinic to the laboratory. *Perspectives on Psychological Science, 17*(4), 1120–1149. https://doi.org/10.1177/17456916211039852

Kredlow, M. A., Eichenbaum, H., & Otto, M. W. (2018). Memory creation and modification: Enhancing the treatment of psychological disorders. *American Psychologist, 73*(3), 269–285. https://doi.org/10.1037/amp0000185

Kuhn, M., Gerlicher, A. M. V., & Lonsdorf, T. B. (2022). Navigating the manyverse of skin conductance response quantification approaches – A direct comparison of trough-to-peak, baseline correction, and model-based approaches in ledalab and PsPM. *Psychophysiology, 59*(9), Article e14058. https://doi.org/10.1111/psyp.14058

Kuznetsova, A., Brockhoff, P. B., & Christensen, R. H. B. (2017). lmerTest package: Tests in linear mixed effects models. *Journal of Statistical Software, 82*, 1–26. https://doi.org/10.18637/jss.v082.i13

Laing, P. A. F., Vervliet, B., Dunsmoor, J. E., & Harrison, B. J. (2025). Pavlovian safety learning: An integrative theoretical review. *Psychonomic Bulletin & Review, 32*, 176–202. https://doi.org/10.3758/s13423-024-02559-4

Lakens, D. (2013). Calculating and reporting effect sizes to facilitate cumulative science: A practical primer for t-tests and ANOVAs. *Frontiers in Psychology, 4*, 863. https://doi.org/10.3389/fpsyg.2013.00863

Lang, P. J., Bradley, M. M., & Cuthbert, B. N. (2008). *International affective picture system (IAPS): Affective ratings of pictures and instruction manual (technical report A-8)*. University of Florida.

Langner, O., Dotsch, R., Bijlstra, G., Wigboldus, D. H. J., Hawk, S. T., & van Knippenberg, A. (2010). Presentation and validation of the radboud faces database. *Cognition & Emotion, 24*(8), 1377–1388. https://doi.org/10.1080/02699930903485076

Le Pelley, M. E. (2004). The role of associative history in models of associative learning: A selective review and a hybrid model. *The Quarterly Journal of Experimental Psychology Section B, 57*(3), 193–243. https://doi.org/10.1080/02724990344000141

LeDoux, J., & Daw, N. D. (2018). Surviving threats: Neural circuit and computational implications of a new taxonomy of defensive behaviour. *Nature Reviews Neuroscience, 19*(5), 269–282. https://doi.org/10.1038/nrn.2018.22

Lenth, R. V., Banfai, B., Bolker, B., Buerkner, P., Giné-Vázquez, I., Herve, M., Jung, M., Love, J., Miguez, F., Piaskowski, J., Riebl, H., & Singmann, H. (2024). Emmeans: Estimated marginal means, Aka least-squares means Version 1.10.5. https://cran.r-project.org/web/packages/emmeans/index.html.

Leuchs, L., Schneider, M., & Spoormaker, V. I. (2019). Measuring the conditioned response: A comparison of pupillometry, skin conductance, and startle electromyography. *Psychophysiology, 56*(1), Article e13283. https://doi.org/10.1111/psyp.13283

Levy, I., & Schiller, D. (2021). Neural computations of threat. *Trends in Cognitive Sciences, 25*(2), 151–171. https://doi.org/10.1016/j.tics.2020.11.007

Li, J., Schiller, D., Schoenbaum, G., Phelps, E. A., & Daw, N. D. (2011). Differential roles of human striatum and amygdala in associative learning. *Nature Neuroscience, 14* (10), 1250–1252. https://doi.org/10.1038/nn.2904

Lissek, S., & van Meurs, B. (2015). Learning models of PTSD: Theoretical accounts and psychobiological evidence. *International Journal of Psychophysiology, 98*(3), 594–605. https://doi.org/10.1016/j.ijpsycho.2014.11.006. Part 2.

Lonsdorf, T. B., Gerlicher, A., Klingelhöfer-Jens, M., & Krypotos, A.-M. (2022). Multiverse analyses in fear conditioning research. *Behaviour Research and Therapy, 153*, Article 104072. https://doi.org/10.1016/j.brat.2022.104072

Lonsdorf, T. B., Klingelhöfer-Jens, M., Andreatta, M., Beckers, T., Chalkia, A., Gerlicher, A., Jentsch, V. L., Meir Drexler, S., Mertens, G., Richter, J., Sjouwerman, R., Wendt, J., & Merz, C. J. (2019). Navigating the garden of forking paths for data exclusions in fear conditioning research. *eLife, 8*, Article e52465. https://doi.org/10.7554/eLife.52465

Lonsdorf, T. B., Menz, M. M., Andreatta, M., Fullana, M. A., Golkar, A., Haaker, J., Heitland, I., Hermann, A., Kuhn, M., Kruse, O., Meir Drexler, S., Meulders, A., Nees, F., Pittig, A., Richter, J., Römer, S., Shiban, Y., Schmitz, A., Straube, B., … Merz, C. J. (2017). Don't fear 'fear conditioning': Methodological considerations for the design and analysis of studies on human fear acquisition, extinction, and return of fear. *Neuroscience & Biobehavioral Reviews, 77*, 247–285. https://doi.org/10.1016/j.neubiorev.2017.02.026

Lonsdorf, T. B., & Merz, C. J. (2017). More than just noise: Inter-individual differences in fear acquisition, extinction and return of fear in humans - Biological, experiential, temperamental factors, and methodological pitfalls. *Neuroscience & Biobehavioral Reviews, 80*, 703–728. https://doi.org/10.1016/j.neubiorev.2017.07.007

Lundqvist, D., Flykt, A., & Öhman, A. (1998). *The Karolinska directed emotional Faces—KDEF*. Karolinska institutet, department of clinical neuroscience. *Psychology Section*.

Mackintosh, N. J. (1975). A theory of attention: Variations in the associability of stimuli with reinforcement. *Psychological Review, 82*(4), 276–298. https://doi.org/10.1037/h0076778

Mair, P., & Wilcox, R. (2020). Robust statistical methods in R using the WRS2 package. *Behavior Research Methods, 52*(2), 464–488. https://doi.org/10.3758/s13428-019-01246-w

McNally, R. J. (1987). Preparedness and phobias: A review. *Psychological Bulletin, 101*(2), 283–303. https://doi.org/10.1037/0033-2909.101.2.283

Milad, M. R., & Quirk, G. J. (2012). Fear extinction as a model for translational neuroscience: Ten years of progress. *Annual Review of Psychology, 63*(1), 129–151. https://doi.org/10.1146/annurev.psych.121208.131631

Mobbs, D., & Kim, J. J. (2015). Neuroethological studies of fear, anxiety, and risky decision-making in rodents and humans. *Current Opinion in Behavioral Sciences, 5*, 8–15. https://doi.org/10.1016/j.cobeha.2015.06.005

Mobbs, D., Wise, T., Suthana, N., Guzmán, N., Kriegeskorte, N., & Leibo, J. Z. (2021). Promises and challenges of human computational ethology. *Neuron, 109*(14), 2224–2238. https://doi.org/10.1016/j.neuron.2021.05.021

Moors, A., Ellsworth, P. C., Scherer, K. R., & Frijda, N. H. (2013). Appraisal theories of emotion: State of the art and future development. *Emotion Review, 5*(2), 119–124. https://doi.org/10.1177/1754073912468165

Ney, L. J., O'Donohue, M. P., Lowe, B. G., & Lipp, O. V. (2022). Angry and fearful compared to happy or neutral faces as conditional stimuli in human fear conditioning: A systematic review and meta-analysis. *Neuroscience & Biobehavioral Reviews, 139*, Article 104756. https://doi.org/10.1016/j.neubiorev.2022.104756

Ney, L. J., Wade, M., Reynolds, A., Zuj, D. V., Dymond, S., Matthews, A., & Felmingham, K. L. (2018). Critical evaluation of current data analysis strategies for psychophysiological measures of fear conditioning and extinction in humans. *International Journal of Psychophysiology, 134*, 95–107. https://doi.org/10.1016/j.ijpsycho.2018.10.010

Niv, Y., Edlund, J. A., Dayan, P., & O'Doherty, J. P. (2012). Neural prediction errors reveal a risk-sensitive reinforcement-learning process in the human brain. *Journal of Neuroscience, 32*(2), 551–562. https://doi.org/10.1523/JNEUROSCI.5498-10.2012

Niv, Y., & Schoenbaum, G. (2008). Dialogues on prediction errors. *Trends in Cognitive Sciences, 12*(7), 265–272. https://doi.org/10.1016/j.tics.2008.03.006

Norbury, A., Brinkman, H., Kowalchyk, M., Monti, E., Pietrzak, R. H., Schiller, D., & Feder, A. (2022). Latent cause inference during extinction learning in trauma-exposed individuals with and without PTSD. *Psychological Medicine, 52*(16), 3834–3845. https://doi.org/10.1017/S0033291721000647

Öhman, A., & Dimberg, U. (1978). Facial expressions as conditioned stimuli for electrodermal responses: A case of "preparedness". *Journal of Personality and Social Psychology, 36*(11), 1251–1258. https://doi.org/10.1037/0022-3514.36.11.1251

Öhman, A., Eriksson, A., & Olofsson, C. (1975). One-trial learning and superior resistance to extinction of autonomic responses conditioned to potentially phobic stimuli. *Journal of Comparative & Physiological Psychology, 88*(2), 619–627. https://doi.org/10.1037/h0078388

Öhman, A., Fredrikson, M., Hugdahl, K., & Rimmo, P.-A. (1976). The premise of equipotentiality in human classical conditioning: Conditioned electrodermal responses to potentially phobic stimuli. *Journal of Experimental Psychology: General, 105*(4), 313–337. https://doi.org/10.1037/0096-3445.105.4.313

Öhman, A., & Mineka, S. (2001). Fears, phobias, and preparedness: Toward an evolved module of fear and fear learning. *Psychological Review, 108*(3), 483–522. https://doi.org/10.1037/0033-295X.108.3.483

Ojala, K. E., & Bach, D. R. (2020). Measuring learning in human classical threat conditioning: Translational, cognitive and methodological considerations. *Neuroscience & Biobehavioral Reviews, 114*, 96–112. https://doi.org/10.1016/j.neubiorev.2020.04.019

Olsson, A., Ebert, J. P., Banaji, M. R., & Phelps, E. A. (2005). The role of social groups in the persistence of learned fear. *Science, 309*(5735), 785–787. https://doi.org/10.1126/science.1113551

Oyarzun, J. P., Kuntz, T. M., Stussi, Y., Karaman, O. T., Vranos, S., Callaghan, B. L., Huttenhower, C., LeDoux, J. E., & Phelps, E. A. (2022). Human threat learning is associated with gut microbiota composition. *PNAS Nexus, 1*(5), Article pgac271. https://doi.org/10.1093/pnasnexus/pgac271

Palminteri, S., Wyart, V., & Koechlin, E. (2017). The importance of falsification in computational cognitive modeling. *Trends in Cognitive Sciences, 21*(6), 425–433. https://doi.org/10.1016/j.tics.2017.03.011

Paskewitz, S., Stoddard, J., & Jones, M. (2022). Explaining the return of fear with revised rescorla-wagner models. *Computational Psychiatry, 6*(1), 213–237. https://doi.org/10.5334/cpsy.88

Pavlov, I. P. (1927). *Conditioned reflexes*. Oxford University Press.

Pearce, J. M., & Hall, G. (1980). A model for Pavlovian learning: Variations in the effectiveness of conditioned but not of unconditioned stimuli. *Psychological Review, 87*(6), 532–552. https://doi.org/10.1037/0033-295X.87.6.532

Pelli, D. G. (1997). The VideoToolbox software for visual psychophysics: Transforming numbers into movies. *Spatial Vision, 10*(4), 437–442. https://doi.org/10.1163/156856897X00366

Phelps, E. A. (2006). Emotion and cognition: Insights from studies of the human amygdala. *Annual Review of Psychology, 57*(1), 27–53. https://doi.org/10.1146/annurev.psych.56.091103.070234

Pool, E. R., Brosch, T., Delplanque, S., & Sander, D. (2016). Attentional bias for positive emotional stimuli: A meta-analytic investigation. *Psychological Bulletin, 142*(1), 79–106. https://doi.org/10.1037/bul0000026

Pool, E. R., Pauli, W. M., Cross, L., & O'Doherty, J. P. (2023). Neural substrates of parallel devaluation-sensitive and devaluation-insensitive Pavlovian learning in humans. *Nature Communications, 14*(1), 8057. https://doi.org/10.1038/s41467-023-43747-5

Posit team. (2024). *RStudio: Integrated development environment for R*. Posit Software, PBC [Computer software] http://www.posit.co/.

R Core Team. (2024). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing [Computer software] https://www.R-project.org/.

Rescorla, R. A. (1988). Pavlovian conditioning: It's not what you think it is. *American Psychologist, 43*(4), 151–160. https://doi.org/10.1037/0003-066X.43.3.151

Rescorla, R. A., & Wagner, A. R. (1972). A theory of Pavlovian conditioning: Variation in the effectiveness of reinforcement and non reinforcement. In A. H. Black, & W. F. Prosaky (Eds.), *Classical conditioning II: Current research and theory* (pp. 64–99). Appleton Century Crofts.

Rigoux, L., Stephan, K. E., Friston, K. J., & Daunizeau, J. (2014). Bayesian model selection for group studies—Revisited. *NeuroImage, 84*, 971–985. https://doi.org/10.1016/j.neuroimage.2013.08.065

Sander, D., Grafman, J., & Zalla, T. (2003). The human amygdala: An evolved system for relevance detection. *Reviews in the Neurosciences, 14*(4), 303–316. https://doi.org/10.1515/REVNEURO.2003.14.4.303

Sander, D., Grandjean, D., & Scherer, K. R. (2018). An appraisal-driven componential approach to the emotional brain. *Emotion Review, 10*(3), 219–231. https://doi.org/10.1177/1754073918765653

Scherer, K. R., & Moors, A. (2019). The emotion process: Event appraisal and component differentiation. *Annual Review of Psychology, 70*(1), 719–745. https://doi.org/10.1146/annurev-psych-122216-011854

Schiller, D., Yu, A. N. C., Alia-Klein, N., Becker, S., Cromwell, H. C., Dolcos, F., Eslinger, P. J., Frewen, P., Kemp, A. H., Pace-Schott, E. F., Raber, J., Silton, R. L., Stefanova, E., Williams, J. H. G., Abe, N., Aghajani, M., Albrecht, F., Alexander, R., Anders, S., … Lowe, L. (2024). The human affectome. *Neuroscience & Biobehavioral Reviews, 158*, Article 105450. https://doi.org/10.1016/j.neubiorev.2023.105450

Schurr, R., Reznik, D., Hillman, H., Bhui, R., & Gershman, S. J. (2024). Dynamic computational phenotyping of human cognition. *Nature Human Behaviour, 8*, 917–931. https://doi.org/10.1038/s41562-024-01814-x

Seligman, M. E. P. (1970). On the generality of the laws of learning. *Psychological Review, 77*(5), 406–418. https://doi.org/10.1037/h0029790

Seligman, M. E. P. (1971). Phobias and preparedness. *Behavior Therapy, 2*(3), 307–320. https://doi.org/10.1016/S0005-7894(71)80064-3

Siddle, D. A. T., & Bond, N. W. (1988). Avoidance learning, Pavlovian conditioning, and the development of phobias. *Biological Psychology, 27*(2), 167–183. https://doi.org/10.1016/0301-0511(88)90048-8

Singmann, H., Bolker, B., Westfall, J., Aust, F., Ben-Shachar, M. S., Højsgaard, S., Fox, J., Lawrence, M. A., Mertens, U., Love, J., Lenth, R., & Christensen, R. H. B. (2024). Afex: Analysis of factorial experiments Version 1.4-1. https://cran.r-project.org/web/packages/afex/index.html.

Society for Psychophysiological Research Ad Hoc Committee on Electrodermal Measures. (2012). Publication recommendations for electrodermal measurements. *Psychophysiology, 49*(8), 1017–1034. https://doi.org/10.1111/j.1469-8986.2012.01384.x

Starita, F., Kroes, M. C. W., Davachi, L., Phelps, E. A., & Dunsmoor, J. E. (2019). Threat learning promotes generalization of episodic memory. *Journal of Experimental Psychology: General, 148*(8), 1426–1434. https://doi.org/10.1037/xge0000551

Starita, F., Stussi, Y., Garofalo, S., & Di Pellegrino, G. (2023). Threat learning in space: How stimulus-outcome spatial compatibility modulates conditioned skin conductance response. *International Journal of Psychophysiology, 190*, 30–41. https://doi.org/10.1016/j.ijpsycho.2023.06.003

Stussi, Y., Brosch, T., & Sander, D. (2015). Learning to fear depends on emotion and gaze interaction: The role of self-relevance in fear learning. *Biological Psychology, 109*, 232–238. https://doi.org/10.1016/j.biopsycho.2015.06.008

Stussi, Y., Ferrero, A., Pourtois, G., & Sander, D. (2019). Achievement motivation modulates Pavlovian aversive conditioning to goal-relevant stimuli. *Npj Science of Learning, 4*(1), 4. https://doi.org/10.1038/s41539-019-0043-3

Stussi, Y., & Pool, E. R. (2022). Multicomponential affective processes modulating food-seeking behaviors. *Current Opinion in Behavioral Sciences, 48*, Article 101226. https://doi.org/10.1016/j.cobeha.2022.101226

Stussi, Y., Pourtois, G., Olsson, A., & Sander, D. (2021). Learning biases to angry and happy faces during Pavlovian aversive conditioning. *Emotion, 21*(4), 742–756. https://doi.org/10.1037/emo0000733

Stussi, Y., Pourtois, G., & Sander, D. (2018). Enhanced Pavlovian aversive conditioning to positive emotional stimuli. *Journal of Experimental Psychology: General, 147*(6), 905–923. https://doi.org/10.1037/xge0000424

Stussi, Y., & Sander, D. (2024). Computational analysis, appraised concern-relevance, and the amygdala: The algorithmic value of appraisal processes in emotion. *Neuroscience & Biobehavioral Reviews, 161*, Article 105676. https://doi.org/10.1016/j.neubiorev.2024.105676

Stussi, Y., Sennwald, V., Pool, E. R., Delplanque, S., Brosch, T., Bianchi-Demicheli, F., & Sander, D. (2021). Individual concerns modulate reward-related learning and behaviors involving sexual outcomes. *Motivation Science, 7*(4), 424–438. https://doi.org/10.1037/mot0000249

Tzovara, A., Korn, C. W., & Bach, D. R. (2018). Human pavlovian fear conditioning conforms to probabilistic learning. *PLoS Computational Biology, 14*(8), Article e1006243. https://doi.org/10.1371/journal.pcbi.1006243

Van Duuren, M., Kendell-Scott, L., & Stark, N. (2003). Early aesthetic choices: Infant preferences for attractive premature infant faces. *International Journal of Behavioral Development, 27*(3), 212–219. https://doi.org/10.1080/01650250244000218

Vrizzi, S., Najar, A., Lemogne, C., Palminteri, S., & Lebreton, M. (2025). Behavioral, computational and self-reported measures of reward and punishment sensitivity as predictors of mental health characteristics. *Nature Mental Health, 3*, 654–666. https://doi.org/10.1038/s44220-025-00427-1.

Westfall, J., Kenny, D. A., & Judd, C. M. (2014). Statistical power and optimal design in experiments in which samples of participants respond to samples of stimuli. *Journal of Experimental Psychology: General, 143*(5), 2020–2045. https://doi.org/10.1037/xge0000014

Wickham, H. (2016). *ggplot2: Elegant graphics for data analysis*. Springer International Publishing. https://doi.org/10.1007/978-3-319-24277-4

Wickham, H., Averick, M., Bryan, J., Chang, W., McGowan, L. D., François, R., Grolemund, G., Hayes, A., Henry, L., Hester, J., Kuhn, M., Pedersen, T. L., Miller, E., Bache, S. M., Müller, K., Ooms, J., Robinson, D., Seidel, D. P., Spinu, V., … Yutani, H. (2019). Welcome to the tidyverse. *Journal of Open Source Software, 4*(43), 1686. https://doi.org/10.21105/joss.01686

Wilson, R. C., & Collins, A. G. (2019). Ten simple rules for the computational modeling of behavioral data. *eLife, 8*, Article e49547. https://doi.org/10.7554/eLife.49547

Wise, T., & Dolan, R. J. (2020). Associations between aversive learning processes and transdiagnostic psychiatric symptoms in a general population sample. *Nature Communications, 11*(1), 4179. https://doi.org/10.1038/s41467-020-17977-w

Wuensch, L., Pool, E. R., & Sander, D. (2021). Individual differences in learning positive affective value. *Current Opinion in Behavioral Sciences, 39*, 19–26. https://doi.org/10.1016/j.cobeha.2020.11.001

Yeo, G. C., & Ong, D. C. (2024). Associations between cognitive appraisals and emotions: A meta-analytic review. *Psychological Bulletin, 150*(12), 1440–1471. https://doi.org/10.1037/bul0000452

Zhang, S., Mano, H., Ganesh, G., Robbins, T., & Seymour, B. (2016). Dissociable learning processes underlie human pain conditioning. *Current Biology, 26*(1), 52–58. https://doi.org/10.1016/j.cub.2015.10.066

Zinbarg, R. E., Williams, A. L., & Mineka, S. (2022). A current learning theory approach to the etiology and course of anxiety and related disorders. *Annual Review of Clinical Psychology, 18*(1), 233–258. https://doi.org/10.1146/annurev-clinpsy-072220-021010

Zuj, D. V., & Norrholm, S. D. (2019). The clinical applications and practical relevance of human conditioning paradigms for posttraumatic stress disorder. *Progress in Neuro-Psychopharmacology and Biological Psychiatry, 88*, 339–351. https://doi.org/10.1016/j.pnpbp.2018.08.014